UNIVERSITAT DE BARCELONA

Decadal climate prediction and predictability for climate services

Carlos Delgado Torres

Decadal climate prediction and predictability for climate services

Carlos Delgado Torres

Directed by Markus G. Donat and Albert Soret Miravet

Ph.D. Thesis

Barcelona 2023

UNIVERSITAT DE BARCELONA

Decadal climate prediction and predictability for climate services

Memoria presentada para optar al grado de doctor por la
Universitat de Barcelona

Programa de doctorado en Física

Autor

Carlos Delgado Torres

Directores

Dr. Markus G. Donat

Dr. Albert Soret Miravet

Tutor

Dr. Joan Bech Rustullet

UNIVERSITAT DE BARCELONA

Carlos D.

Decadal climate prediction and predictability for climate services

Ph.D. Thesis
Barcelona, 2023

Author: **Carlos Delgado Torres**
Directors: Markus G. Donat and Albert Soret Miravet

Cover design: Diana Urquiza

Declaration

I confirm that the work submitted in this Ph.D. thesis is my own, except where work which has formed part of jointly-authored publications has been included. I also confirm that appropriate credit has been given within this Ph.D. thesis where reference has been made to the work of others.

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

The research done during the Ph.D. has been developed at the Earth Sciences Department of the Barcelona Supercomputing Center - Centro Nacional de Supercomputación (BSC-CNS). This Ph.D. thesis has received financial support from the Formación de Personal Investigador (**FPI**) program of the Spanish Ministry for Science and Innovation (grant PRE2019-509 08864 financed by MCIN/AEI/10.13039/501100011033 and by FSE invierte en tu futuro). The author is also grateful for productive discussions and engagement with colleagues from other research centres within the context of the **CLINSA** project of the Spanish Ministry for Science, Innovation and Universities (grant CGL2017-85791-R), the **C3S_34c** contract (ECMWF/COPERNICUS/2019/C3S_34c_DWD) of the Copernicus Climate Change Service (C3S) operated by the European Centre for Medium-Range Weather Forecasts (ECMWF), the European Commission H2020 **EUCP** project (grant 776613), the European Commission H2020 **FOCUS-Africa** project (grant 869575) and the Horizon Europe **ASPECT** project (grant 101081460), which allowed collaborations with researchers from the leading scientific community in Europe.

Acknowledgements

I want to express my deep gratitude to the following people whose support and guidance have been essential in the completion of this PhD thesis:

First and foremost, I would like to thank my Ph.D. directors, Markus and Albert, for their invaluable mentorship, continuous encouragement, and insightful feedback throughout this research journey. Thanks a lot for all the scientific discussions since we started working on the Master's thesis long time ago. I appreciate your motivating words during all these years, but especially your reassuring words when things went not so well. I am also profoundly grateful to Paco for his unlimited patience when explaining anything and serving as a source of knowledge. I could not count all the things I have learnt from you all.

Special thanks to Núria, An-Chi and Bala for all your technical and scientific support. It is a pleasure working with you. I also want to thank Marga and Pierre-Antoine for downloading and formatting the huge amount of data we have been analysing. The work done during these years would not have been possible without you.

I want to say special thanks to Cristina, Jaume and Lluís for all the good moments we have had at home. I came to the flat without knowing you, and we had to pass the Covid-19 lockdown together after only one month. I could not have imagined how easy it was going to live with people you have just met. I also want to thank Ángel and Carmen for hosting me during my research stay in New York. Your hospitality and support enhanced my experience in such a chaotic city.

To the Earth System Services group, I express my gratitude for providing a stimulating and collaborative office environment. Huge thanks to several

Climate Services team members: Bala, Núria, Raül, Nadia, Nube, Lluís, Pep, Sara, Alba, Victòria, Aleks, Albert, Jaume, Vero, Eren, Kat and Matias; and also to the Knowledge and Integration team members, especially to Eulàlia, Ilaria, Andria, Gert, Sara, Dragana, Paula and Diana. The good atmosphere in the office makes the routine of going to work the best plan.

I also want to thank the members of the Climate Variability and Change group for their involvement in countless scientific discussions that broadened my horizons and contributed significantly to this work, especially to Roberto, Aude, Vero, Paloma, Andrea, Bianca, Rubén, Rashed, Yohan, Pablo and Paolo. I am also grateful to Cristina, Raquel, Elisa and Dani for all the enjoyable moments, and to Albert Vila, Elisenda and Katia for all the help and support.

Finally, I want to thank my parents and friends in Madrid for their unlimited support and for receiving me every time I returned as if I had never left. Thanks a lot to all of you.

Abstract

Climate variations at annual to decadal time scales affect many regions around the globe, causing direct impacts on the economy, ecosystems and society in several sectors, such as renewable energy, agriculture, food security, water management, fisheries, health, insurance and urban planning. Knowing these variations ahead of time allows for implementing measures to adapt, mitigate and build resilience to the consequences of a changing climate. At annual to decadal time scales, climate variations are caused by both externally generated forcings (which can be natural, e.g. volcanic eruptions and solar radiation, or anthropogenic, e.g. greenhouse gases emissions) and internal climate variability (which is generated by interactions between different components of the climate system and triggers slow, natural oscillations that are connected to regional climate conditions).

Climate models were developed as tools that aim to assess climate dynamics and anticipate climate variations by solving the physical equations that govern the climate system. Climate projections incorporate external forcing information based on different socio-economic scenarios to project possible pathways the climate system would follow. The same models are used to predict climate variability and change at annual to decadal time scales by also incorporating information on the current climate state. This is done through a model initialisation process in which observation-based data, referred to as initial conditions, is included in the models to phase the modelled simulations with the observed climate state.

However, climate predictions do not necessarily capture climate variations correctly, as model inefficiencies, errors in the initial conditions and mathematical approximations degrade the forecast quality, thus limiting the potential usefulness of the predictions for decision-makers. Besides, not all

variations might be predictable due to chaotic characteristics of the climate system. Therefore, the forecast quality assessment is an essential step prior to using any climate information derived from models to ensure that such information is trustable and beneficial for decision-making. Forecast quality assessment involves comparing the climate hindcasts (i.e. retrospective climate predictions) to past observations to evaluate their degree of agreement, thus having an estimate of how the models could perform in predicting future climate variations.

In addition to evaluating the forecast quality, climate hindcasts also allow for applying post-processing techniques to the predictions in order to correct their systematic biases through calibration techniques (e.g. the modelled and observed climatologies may differ) or downscale the climate information to provide regional information (as the original spatial resolution may be too coarse for regional decisions), among others. Besides, estimating the quality of different models (or a combination of them, i.e. a multi-model ensemble) allows for selecting the best climate information for each specific variable, region and forecast period.

The work developed within this Ph.D. thesis has focused on the evaluation of the forecast quality for predictions of several variables, indices and indicators relevant for decision-making in several sectors, with a particular focus on agriculture. The evaluation has been performed globally, for the individual models and multi-model ensemble, and different forecast periods in order to identify windows of opportunity for which the climate predictions show enough quality to be used for decision-making. Furthermore, the historical forcing simulations (i.e. retrospective climate projections) can be considered an alternative prediction system that simulates climate in response to forcing changes, and have been included in all analyses to estimate the impact of model initialisation and find the best source of climate information for climate variations during the next years.

First, the quality of multi-model forecasts of temperature, precipitation, the Atlantic Multidecadal Variability index (AMV) and Global near-Surface Air

Temperature anomalies (GSAT) generated from all decadal hindcasts contributing to the Coupled Model Intercomparison Project Phase 6 (CMIP6) were evaluated, finding high skill for predictions of temperature, AMV and GSAT, and limited skill for predictions of precipitation. The skill of several approaches for building a multi-model ensemble were compared, identifying only small differences between them, and the impact of calibration on the predictions. The multi-model ensemble was also compared to individual forecast systems, finding that the best system for each particular location usually outperforms the multi-model ensemble. However, the multi-model provides higher skill than at least half of the systems. Therefore, the best system can be selected to provide the highest-quality climate information. However, the multi-model ensemble is the best choice for a systematic forecast provision where several variables, forecast periods and regions are included. The decadal multi-model ensemble was also compared to the CMIP6 historical simulations multi-model, finding an added value from model initialisation over several ocean and land regions for temperature, and for AMV and GSAT. Given the low availability of timely decadal forecasts in near-real time, the full multi-model was compared to a sub-ensemble of predictions generated from forecast systems that provided timely forecasts to assess the impact of the ensemble size in an operational climate services context.

Second, the representation and prediction of the Euro-Atlantic weather regimes by the EC-Earth3 model were assessed identifying the dominating atmospheric circulation patterns in this region by applying the k-means clustering algorithm to daily fields of sea level pressure. Skilful predictions of weather regimes could be used as a source of predictability for local climate conditions, and thus translated into useful climate information for decision-making. The Euro-Atlantic weather regimes are the positive and negative phases of the North Atlantic Oscillation (NAO+ and NAO-, respectively), Blocking, and Atlantic Ridge in winter; and the NAO-, Blocking, Atlantic Ridge, and Atlantic Low in summer. The EC-Earth3 correctly represents the spatial patterns and climatological frequencies of all weather regimes. However, the skill in predicting the annual to decadal variations

of the weather regimes' frequency of occurrence is low, and the model initialisation does not improve such prediction skill.

Third, the multi-model forecast quality of the CMIP6 decadal hindcasts is evaluated for multi-annual predictions of a set of indices related to the frequency and intensity of daily temperature and precipitation extremes. Such predictions are essential to develop adaptation plans and anticipate the impacts of extreme events ahead of time in several climate-sensitive sectors. The multi-model ensemble is skillful in predicting temperature extremes over most land regions, while the quality is lower for precipitation extremes. Comparing the skill with that for mean temperature and precipitation, extremes are predicted with lower skill, especially those related to the most extreme days. Compared to the historical forcing simulations, decadal predictions show only small and region-dependent skill improvements from model initialisation.

Finally, this Ph.D. thesis presents the applications of the research within several European projects and a contract with a private company for which prototypes of climate services have been created. For instance, prototypes of forecast products have been developed for the Southern African Development Community, Tanzania and Malawi regions. These prototypes consist of annual and multi-annual forecast products of temperature, precipitation and drought conditions. Another application was the development of forecast products of climate variables relevant to cotton production. In this case, only the crop months for each location were considered when creating the forecasts to tailor the predictions to that season of the year.

Resumen

Las variaciones climáticas en escalas temporales de uno a diez años afectan a muchas regiones del globo, causando impactos directos en la economía, los ecosistemas y la sociedad en varios sectores, como el de las energías renovables, la agricultura, la seguridad alimentaria, la gestión del agua, la pesca, la sanidad, los seguros y la planificación urbana. Conocer estas variaciones climáticas con antelación permite implementar medidas de adaptación, mitigación y resiliencia a las consecuencias de un clima variable. En escalas de tiempo anual a decadal, las variaciones climáticas están causadas tanto por forzamientos generados externamente (que pueden ser naturales, como las erupciones volcánicas y la radiación solar, o antropogénicos, como las emisiones de gases de efecto invernadero) como por la variabilidad climática interna (que se genera por las interacciones entre los distintos componentes del sistema climático y desencadena oscilaciones lentas y naturales que se relacionan con condiciones climáticas a escala regional).

Los modelos climáticos se desarrollaron como herramientas para estudiar la dinámica climática y anticipar las variaciones del clima resolviendo las ecuaciones físicas que rigen el sistema climático. Las proyecciones climáticas incorporan información sobre forzamientos externos basada en distintos escenarios socioeconómicos para proyectar posibles trayectorias que seguiría el sistema climático. Los mismos modelos se utilizan para predecir la variabilidad y el cambio climático a escalas temporales anuales y decadales, incorporando también información sobre el estado actual del clima. Este proceso se conoce como inicialización del modelo, en el que las condiciones iniciales (información basada en datos observados) se incluye en el modelo para ajustar la fase de variabilidad con el estado del clima observado.

Sin embargo, las predicciones climáticas no siempre predicen correctamente las variaciones climáticas, ya que deficiencias en los modelos, errores en las condiciones iniciales y aproximaciones matemáticas degradan la calidad de las predicciones, lo que limita su utilidad para los usuarios. Además, no todas las variaciones son predecibles debido a la naturaleza caótica del sistema climático. Por lo tanto, la evaluación de la calidad de las predicciones es un paso esencial antes de utilizar cualquier información climática derivada de modelos para garantizar que dicha información sea fiable y beneficiosa para la toma de decisiones. La evaluación de la calidad de las previsiones consiste en comparar las predicciones del pasado con las observaciones para evaluar su grado de concordancia y, de este modo, tener una estimación de cómo podrían funcionar los modelos para predecir futuras variaciones climáticas.

Además de evaluar la calidad, las predicciones del pasado también permiten aplicar técnicas de postprocesado a las predicciones del futuro para corregir errores sistemáticos mediante técnicas de calibración (por ejemplo, las climatologías del modelo y observadas pueden ser diferentes) o incrementar la resolución espacial de las predicciones para proporcionar información regional (ya que la resolución espacial original puede ser demasiado gruesa para la toma de decisiones regionales), entre otras. Además, evaluar la calidad de diferentes modelos (o una combinación de ellos, es decir, de un multimodelo) permite seleccionar la mejor fuente de información climática para cada variable, región y periodo de predicción.

El trabajo desarrollado en esta tesis doctoral se centra en la evaluación de la calidad de las predicciones climáticas de diversas variables, índices e indicadores relevantes para la toma de decisiones en diversos sectores, con especial atención a la agricultura. La evaluación se ha realizado de forma global, para los modelos individuales y el multimodelo, y diferentes períodos de predicción con el fin de identificar los casos en los que las predicciones climáticas tienen suficiente calidad como para ser utilizadas en la toma de decisiones. Además, las simulaciones de forzamiento histórico (es decir, las proyecciones climáticas del pasado) pueden considerarse un

sistema de predicción alternativo que simula el clima en respuesta a los cambios de los forzamientos externos, y se han incluido en todos los análisis para estimar el impacto de la inicialización del modelo y encontrar la mejor fuente de información climática para las variaciones climáticas para los próximos años.

En primer lugar, se evaluó la calidad de las predicciones multimodelo de temperatura, precipitación, índice de la Variabilidad Multidecadal de Atlántico (AMV) y anomalía de la Temperatura Global del Aire en Superficie (GSAT) generadas a partir de todos las predicciones decadales que contribuyen a la Fase 6 del Proyecto de Intercomparación de Modelos Acoplados (CMIP6). En este estudio se encontró una calidad alta para las predicciones de temperatura, AMV y GSAT, mientras que la calidad es más limitada para las predicciones de precipitación. En el estudio se comparó la calidad de diferentes modos de construir un multimodelo, identificando sólo pequeñas diferencias entre ellos, y el efecto de la calibración en las predicciones. También se comparó la calidad del multimodelo contra los modelos individuales, y se observó que el mejor modelo para cada lugar concreto suele ser mejor que el multimodelo. Sin embargo, el multimodelo proporciona una mayor precisión que, al menos, la mitad de los sistemas. Por lo tanto, se puede seleccionar el mejor modelo para proporcionar la información climática de mayor calidad para cada caso particular. Sin embargo, el multimodelo es la mejor opción para una predicción sistemática en la que se incluyen varias variables, períodos de predicción y regiones. El multimodelo decadal también se comparó con el multimodelo de simulaciones históricas, encontrando un beneficio de la inicialización del modelo en varias regiones oceánicas y terrestres para temperatura, y para la AMV y GSAT. Dada la escasa disponibilidad de predicciones decadales en tiempo real, se comparó el multimodelo completo con un subconjunto de predicciones generadas con los modelos que proporcionan predicciones en tiempo real para evaluar el impacto del tamaño del multimodelo en un contexto de servicios climáticos operativos.

En segundo lugar, se evaluó la representación y predicción de los tipos de tiempo euro-atlánticos por el modelo EC-Earth3, identificando los patrones espaciales dominantes de circulación atmosférica en esta región mediante la aplicación del algoritmo de agrupación k-means sobre campos diarios de presión a nivel del mar. Las predicciones de alta calidad de los tipos de tiempo podrían utilizarse como fuente de predictibilidad para las condiciones climáticas locales, y traducirse así en información climática útil para la toma de decisiones. Los tipos de tiempo euro-atlánticos son las fases positiva y negativa de la Oscilación del Atlántico Norte (NAO+ y NAO-, respectivamente), el Bloqueo y la Dorsal Atlántica en invierno; y la NAO-, el Bloqueo, la Dorsal Atlántica y la Baja Atlántica en verano. Se vio que el EC-Earth3 representa correctamente los patrones espaciales y las frecuencias climatológicas de todos los tipos de tiempo. Sin embargo, la habilidad para predecir las variaciones anuales a decadales de la frecuencia de ocurrencia de los tipos de tiempo es baja, y la inicialización del modelo no mejora dicha habilidad de predicción.

En tercer lugar, se evalúa la calidad de las predicciones decadales multimodelo para predicciones multianuales de un conjunto de índices relacionados con la frecuencia e intensidad de extremos de temperaturas y precipitaciones. Estas predicciones son esenciales para elaborar planes de adaptación y anticiparse a los efectos de los eventos extremos en varios sectores sensibles al clima. La calidad del multimodelo es alta para la predicción de eventos extremos de temperatura en la mayoría de las regiones terrestres, mientras que la calidad es menor para el caso de precipitación extrema. Comparando dicha calidad contra la correspondiente para predicciones de temperatura y la precipitación media, los extremos se predicen con menor calidad, especialmente los relacionados con los días más extremos. En comparación con las simulaciones de forzamiento histórico, las predicciones decadales muestran generalmente poco beneficio gracias a la inicialización del modelo, el cuál depende de la región de interés.

Por último, esta tesis doctoral presenta las aplicaciones de la investigación desarrollada en el marco de varios proyectos europeos y contratos con empresas, para los cuales se han creado prototipos de servicios climáticos. Por ejemplo, se han creado prototipos de servicios climáticos para las regiones de la Comunidad para el Desarrollo del África Meridional, Tanzania y Malawi. Estos prototipos consisten en productos que contienen predicciones anuales y multianuales de temperatura, precipitación y condiciones de sequía. Otra aplicación ha sido el desarrollo de productos con información de variables climáticas relevantes para la producción de algodón. En este caso, a la hora de crear la información climática, sólo se consideraron los meses de cosecha de cada lugar para adaptar las predicciones a esa época del año.

Resum[*]

Les variacions climàtiques en escales temporals d'un a deu anys afecten moltes regions del mon, causant impactes directes a l'economia, els ecosistemes i la societat en diversos sectors, com el de les energies renovables, l'agricultura, la seguretat alimentària, la gestió de l'aigua, la pesca, la sanitat, les assegurances i la planificació urbana. Conèixer aquestes variacions climàtiques amb antelació permet implementar mesures d'adaptació, mitigació i resiliència que facin front a les conseqüències d'un clima canviant. En les escales temporals anuals fins a decadals, les variacions climàtiques estan causades tant per forçaments externs (que poden ser naturals, com les erupcions volcàniques i la radiació solar, o antropogènics, com les emissions de gasos d'efecte hivernacle), com per la variabilitat climàtica interna (que es genera per les interaccions entre els diferents components del sistema climàtic i desencadena oscil·lacions lentes i naturals que es relacionen amb condicions climàtiques a escala regional).

Els models climàtics es van desenvolupar com a eines per estudiar la dinàmica climàtica i anticipar les variacions del clima resolent les equacions físiques que regeixen el sistema climàtic. Les projeccions climàtiques incorporen informació sobre forçaments externs basades en diferents escenaris socioeconòmics per projectar les possibles trajectòries que seguiria el sistema climàtic. Aquests mateixos models es fan servir per predir la variabilitat i el canvi climàtics a escales temporals anuals i decadals, incorporant també informació sobre l'estat actual del clima. Aquest procés es coneix com a inicialització del model, en què les condicions inicials (informació basada en observacions) s'inclouen al model per ajustar-lo amb l'estat del clima observat.

[*] Translated by Eulàlia Baulenas Serra

Tot i així, les prediccions climàtiques no sempre prediuen correctament les variacions climàtiques degut a deficiències en els models, errors en les condicions inicials i aproximacions matemàtiques que poden degradar la qualitat de les prediccions, cosa que limita la seva utilitat per als usuaris. A més a més, no totes les variacions són predictibles a causa de la naturalesa caòtica del sistema climàtic. Per tant, l'avaluació de la qualitat de les prediccions és un pas essencial abans de fer servir qualsevol informació climàtica derivada de models per garantir que aquesta informació sigui fiable i beneficiosa per a la presa de decisions. L'avaluació de la qualitat de les previsions consisteix en comparar les prediccions del passat amb les observacions per avaluar-ne el grau de concordança i, així, tenir una estimació de com podrien funcionar els models per predir futures variacions climàtiques.

A més d'avaluar la qualitat, les prediccions del passat també permeten aplicar tècniques de postprocessament a les prediccions del futur per corregir errors sistemàtics mitjançant tècniques de calibratge (per exemple, les climatologies del model i les observacions poden ser diferents) o incrementar la resolució espacial de les prediccions per proporcionar informació regional (ja que la resolució espacial original pot ser massa poc definida per a la presa de decisions regionals), entre d'altres. A més, avaluar la qualitat de diferents models (o una combinació, és a dir, d'un multimodel) permet seleccionar la millor font d'informació climàtica per a cada variable, regió i període de predicció.

El treball desenvolupat en aquesta tesi doctoral se centra en l'avaluació de la qualitat de les prediccions climàtiques de diverses variables, índexs i indicadors rellevants per a la presa de decisions a diversos sectors, amb especial atenció a l'agricultura. L'avaluació s'ha realitzat de forma global, per als models individuals i el multimodel, i diferents períodes de predicció, per tal d'identificar els casos en què les prediccions climàtiques tenen prou qualitat per a ser utilitzades en la presa de decisions. A més, les simulacions de forçament històric (és a dir, les projeccions climàtiques del passat) es poden considerar un sistema de predicció alternatiu que simula el clima en resposta als canvis dels forçaments externs, i s'han inclòs a totes les anàlisis per estimar l'impacte de la inicialització del model i trobar la millor font d'informació climàtica per a les variacions climàtiques dels propers anys.

En primer lloc, s'ha avaluat la qualitat de les prediccions multimodel de temperatura, precipitació, índex de la Variabilitat Multidecadal de l'Atlàntic (AMV) i l'anomalia de la Temperatura Global de l'Aire en Superfície (GSAT) generades a partir de totes les prediccions decadals que contribueixen a la Fase 6 del Projecte d'Intercomparació de Models Acoblats (CMIP6). En aquest estudi es va observar una alta qualitat en les prediccions de temperatura, AMV i GSAT, mentre que la qualitat és més limitada per a les prediccions de precipitació. A l'estudi es va comparar la qualitat de diferents maneres de construir un multimodel, identificant petites diferències entre ells, i l'efecte del calibratge en les prediccions. També es va comparar la qualitat del multimodel contra els models individuals, i es va observar que el millor model per a cada lloc concret sol ser millor que el multimodel. No obstant això, el multimodel proporciona més precisió que, almenys, la meitat dels sistemes. Per tant, es pot seleccionar el millor model per proporcionar la informació climàtica de més qualitat per a cada cas particular. Tot i això, el multimodel és la millor opció per a una predicció sistemàtica en què s'inclouen diverses variables, períodes de predicció i regions. El multimodel decadal també es va comparar amb el multimodel de simulacions històriques, trobant-se un benefici de la inicialització del model a diverses regions oceàniques i terrestres per a temperatura, i per a l'AMV i GSAT. Atesa la poca disponibilitat de prediccions decadals en temps real, es va comparar el multimodel complet amb un subconjunt de prediccions generades amb els models que proporcionaven prediccions en temps real, per tal d'avaluar l'impacte de la mida del multimodel en un context de serveis climàtics operatius.

En segon lloc, s'ha avaluat la representació i la predicció dels tipus de temps euroatlàntics pel model EC-Earth3, identificant els patrons espacials dominants de circulació atmosfèrica en aquesta regió mitjançant l'aplicació de l'algorisme d'agrupació k-means sobre camps diaris de pressió a nivell del mar. Les prediccions d'alta qualitat dels tipus de temps es podrien utilitzar com a font de predictibilitat per a les condicions climàtiques locals, i traduir-se així en informació climàtica útil per a la presa de decisions. Els tipus de temps euroatlàntics es refereixen a les fases positiva i negativa de l'Oscil·lació de l'Atlàntic Nord (NAO+ i NAO-, respectivament), el Bloqueig i la Dorsal Atlàntica a l'hivern; i la NAO-, el Bloqueig, la Dorsal Atlàntica i la Baixa Atlàntica a l'estiu. Es va veure que l'EC-Earth3 representa correctament els patrons espacials i les freqüències climatològiques de tot tipus de temps. Tot i així, l'habilitat per predir les variacions anuals a decadals en la freqüència d'ocurrència dels tipus de temps és baixa, i la inicialització del model no en millora l'habilitat de predicció.

En tercer lloc, s'ha avaluat la qualitat de les prediccions decadals multimodel per a prediccions multianuals d'un conjunt d'índexs relacionats amb la freqüència i la intensitat de temperatures extremes i precipitacions. Aquestes prediccions són essencials per elaborar plans d'adaptació i anticipar-se als efectes dels esdeveniments extrems en diversos sectors sensibles al clima. Els resultats mostren com la qualitat del multimodel és alta per a la predicció de temperatures extremes a la majoria de les regions terrestres, mentre que la qualitat és menor per al cas de precipitació extrema. Comparant aquesta qualitat contra la corresponent per a prediccions de temperatura i precipitació mitjana, els extrems es prediuen amb menor qualitat, especialment els relacionats amb els dies més extrems. En comparació amb les simulacions de forçament històric, les prediccions decadals mostren generalment poc benefici degut a la inicialització del model, que es mostra depenent de la regió d'interès.

Per acabar, aquesta tesi doctoral presenta aplicacions d'aquesta recerca desenvolupades en el marc de diversos projectes europeus i contractes amb empreses, per als quals s'han creat prototips de serveis climàtics. Per exemple, s'han creat prototips de productes per a les regions de la Comunitat per al Desenvolupament de l'Àfrica Meridional, Tanzània i Malawi. Aquests prototips consisteixen en productes que contenen prediccions anuals i multianuals de temperatura, precipitació i condicions de sequera. Una altra aplicació ha estat el desenvolupament de productes que contenen informació de variables climàtiques rellevants per a la producció de cotó. En aquest cas, a l'hora de crear la informació climàtica només es van considerar els mesos de collita de cada lloc per adaptar les prediccions a aquesta època de l'any.

Abbreviations

- ACC: Anomaly Correlation Coefficient
- AMV: Atlantic Multidecadal Variability
- AO: Arctic Oscillation
- AL: Atlantic Low
- AR: Atlantic Ridge
- BEST: Berkeley Earth Surface Temperatures
- BL: Blocking
- BS: Brier Score
- BSS: Brier Skill Score
- BSC-CNS: Barcelona Supercomputing Center - Centro Nacional de Supercomputación
- C3S: Copernicus Climate Change Service
- CDD: Consecutive Dry Days
- CMCC: Centro Euro-Mediterraneo per i Cambiamenti Climatici
- CMIP5: Coupled Model Intercomparison Project Phase 5
- CMIP6: Coupled Model Intercomparison Project Phase 6
- CRAN: Comprehensive R Archive Network
- CRPSS: Continuous Ranked Probability Skill Score
- CVC: Climate Variability and Change
- CWD: Consecutive Wet Days
- DCPP: Decadal Climate Prediction Project

- DePreSys: Met Office Decadal Prediction System
- DJF: December-January-February
- DWD: Deutscher Wetterdienst
- ECMWF: European Centre for Medium-Range Weather Forecasts
- ENSO: El Niño - Southern Oscillation
- ERA5: 5th generation ECMWF Atmospheric Reanalysis
- ESGF: Earth System Grid Federation
- ES: Earth Sciences
- ESS: Earth System Services
- ETCCDI: Expert Team on Climate Change Detection and Indices
- FDR: False Detection Rate
- FPI: Formación de Personal Investigador
- GHCNv4: Global Historical Climatology Network version 4
- GPCC: Global Precipitation Climatology Project
- GSAT: Global near-Surface Air Temperature
- HIST: Historical forcing simulations
- HPC: High Performance Computing
- IOD: Indian Ocean Dipole
- IPO: Interdecadal Pacific Oscillation
- IRI: International Research Institute for Climate and Society
- JJA: June-July-August
- JRA-55: Japanese 55-year Reanalysis
- LOESS: Locally Estimated Scatterplot Smoothing
- MiKlip: German Mid-term Climate Forecast
- MJO: Madden-Julian Oscillation
- NAO: North Atlantic Oscillation

- NCEP1: NCEP/NCAR Reanalysis 1
- PDO: Pacific Decadal Oscillation
- QBO: Quasi-Biennial Oscillation
- PR: Precipitation
- R90p: Sum of precipitation in days where daily PR exceeds the 95th percentile of daily precipitation
- REGEN: Rainfall Estimates on a Gridded Network
- SADC: Southern African Development Community
- RMSE: Root Mean Squared Error
- RMSSS: Root Mean Squared error Skill Score
- RPS: Ranked Probability Score
- RPSS: Ranked Probability Skill Score
- ROC: Relative Operating Characteristic
- ROCSS: Relative Operating Characteristic Skill Score
- Rx5day: Maximum 5-day consecutive precipitation
- SAM: Southern Annular Mode
- SPEI: Standardised Precipitation Evapotranspiration Index
- SPOD: South Pacific Ocean Dipole
- SSPs: Shared Socioeconomic Pathways
- SST: Sea Surface Temperatures
- SUNSET: SUbseasoNal to decadal climate forecast post-processIng and asSEssmenT
- TAS: Near-Surface Air Temperature
- TN10p: Percentage of days when minimum temperature is below the 10th daily percentile
- TNn: Minimum of daily minimum temperature

- TX90p: Percentage of days when maximum temperature is above the 90th daily percentile
- TXx: Maximum of daily maximum temperature
- UB: Universitat de Barcelona
- United Kingdom's Met Office
- WMO: World Meteorological Organization

Table of contents

Chapter 1

Introduction

The variability of the climate system is modulated by both natural and anthropogenic factors. However, Earth's climate has been experiencing unprecedented changes over the past decades, with the global surface temperature for the decade 2011-2020 reaching values 1.09ºC warmer than the 1850-1900 average and associated with more frequent and intense extreme events such as heatwaves and droughts (IPCC, 2023b). Therefore, the impact of climate change compromises the security of natural ecosystems, human health and socio-economics (Abbass *et al.*, 2022; Pecl *et al.*, 2017).

Decadal climate predictions have been recently developed as a relatively new source of climate information to understand and predict climate evolution for the next few years up to one decade. This annual-to-decadal climate information is essential to provide valuable information for policy and decision-making in several climate-vulnerable sectors such as agriculture, renewable energy, health and water management. High-quality climate predictions allow the development of adaptation and mitigation measures and early-warning systems for climate variations (Curtis *et al.*, 2017; Hanlon *et al.*, 2013; Kushnir *et al.*, 2019; Sillmann *et al.*, 2017).

1.1. Climate system and its variability and changes

The Earth's climate system is formed by several components that interact between them. These components are the atmosphere, hydrosphere, lithosphere, biosphere and cryosphere, and they play different roles in the climate system (Rodó & Comín, 2003). A schematic figure on the different climate system's components, their roles

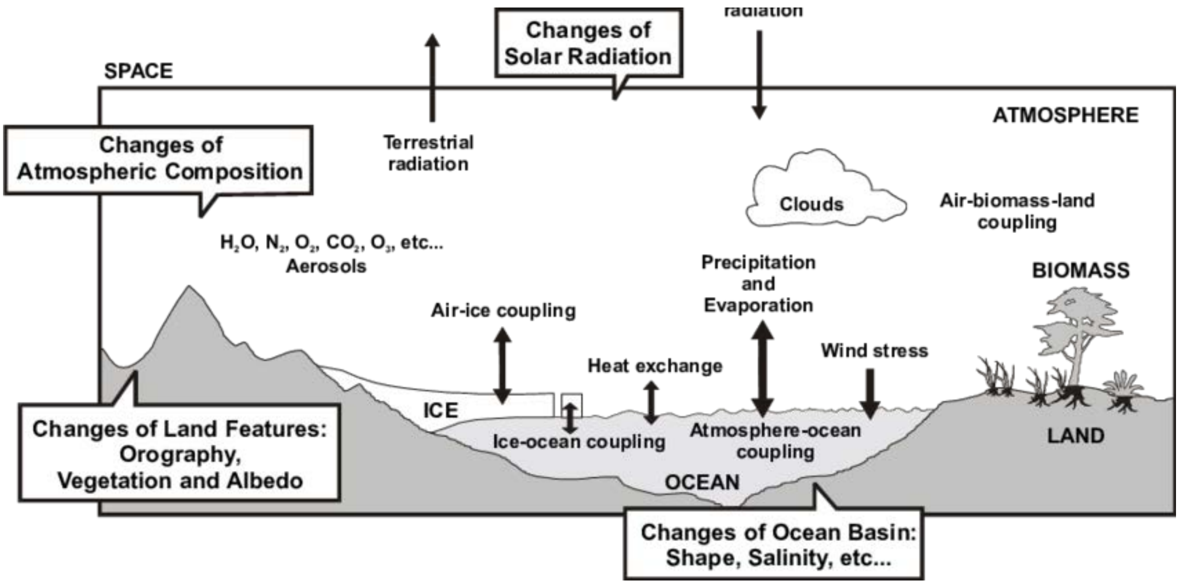

radiation
Changes of Solar Radiation
SPACE
Terrestrial radiation
ATMOSPHERE
Changes of Atmospheric Composition
Clouds
Air-biomass-land coupling
H_2O, N_2, O_2, CO_2, O_3, etc...
Aerosols
Precipitation and Evaporation
BIOMASS
Air-ice coupling
Heat exchange
Wind stress
ICE
Ice-ocean coupling
Atmosphere-ocean coupling
LAND
Changes of Land Features: Orography, Vegetation and Albedo
OCEAN
Changes of Ocean Basin: Shape, Salinity, etc...

Figure 1.1: Schematic representation of the different climate system's components, their roles and interactions. Figure taken from Rodó & Comín (2003).

and interactions can be seen in Figure 1.1. For example, the atmosphere absorbs and reflects terrestrial and solar radiation, helping to distribute the heat around the planet. Weather processes, such as precipitation and wind, take place in the atmosphere. The hydrosphere, mainly formed by the oceans, also plays a role in the distribution of the heat around the planet, as it stores heat from solar radiation, and the currents distribute it around the planet. In addition, the ocean absorbs carbon dioxide from the atmosphere, thus playing a key role in the carbon cycle. The lithosphere and biosphere, which include the land and vegetation, respectively, influence the rest of the climate system depending on the vegetation and land type and use. Finally, the cryosphere helps to regulate the global temperature as the ice reflects a large amount of solar radiation.

The variability of the climate system is modulated by both internal variability and external forcings. Internal variability refers to natural variations produced by the interactions between the different climate system components (such as the ocean-atmosphere dynamic and thermodynamic interactions; Deser *et al.*, 2012) at different time scales. On the other hand, external forcings refer to those factors that affect the climate system without being directly part of it, and can be natural (e.g. volcanic eruptions and variability of the solar radiation) or anthropogenic (e.g. greenhouse gas emissions, land use and deforestation) (Smith *et al.*, 2016).

The interactions between the different climate system's components and the contributions from externally-forced factors result in modes of variability of the climate system. Each mode of variability operates at a particular time scale and area of the globe (Figure 1.2), and provides some predictability due to their low-frequency variations. Thus, they are relevant on specific time scales and regions, and are associated with certain large-scale climate patterns and local weather conditions, such as extreme climate events (Kenyon & Hegerl, 2008). Besides, these modes of variability can also interact with each other (L'Heureux *et al.*, 2017; Trascasa-Castro *et al.*, 2021) and with external forcings (Hermanson *et al.*, 2020), leading to more complex and unpredictable climate patterns and weather conditions.

Some of the main modes of internal variability at subseasonal, seasonal and decadal time scales are the following:

- Madden-Julian Oscillation (MJO; Woolnough *et al.*, 2007): It operates over the Indian and Pacific Oceans, and is characterised by an eastward displacement of tropical rainfall. It consists of eight phases (lasting a total of 1 to 3 months), starting with enhanced convection over the western Indian Ocean which is slowly displaced towards the Central Pacific Ocean, where the enhanced rainfall is dissolved, and the cycle starts again over the Indian Ocean.

- North Atlantic Oscillation (NAO; Visbeck *et al.*, 2001): It is characterised by changes in the atmospheric pressure difference between the Azores High and the Iceland Low. It is related to local weather in Europe, North America and parts of Africa. For instance, its positive phase is related to stronger westerly winds and storm tracks across the North Atlantic, while its negative phase is associated with weaker winds and a shift of the storm tracks towards southern Europe. This shift of the storm tracks causes changes in temperature, precipitation and extreme events, particularly over Europe.

- Arctic Oscillation (AO; Thompson & Wallace, 1998): It is characterised by changes in the atmospheric pressure difference between the Arctic region and mid-latitudes, influencing the position and strength of the polar jet stream. Thus,

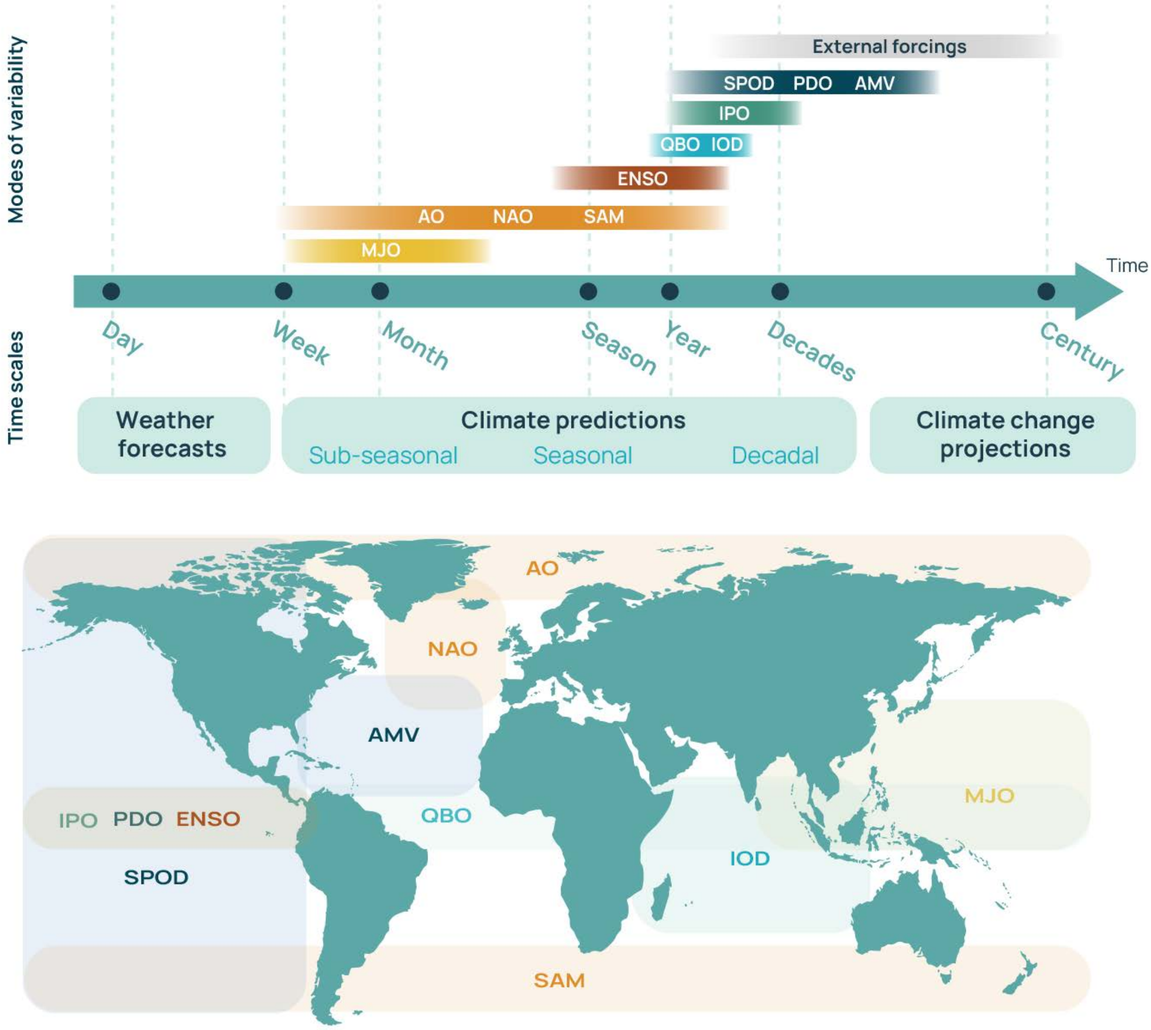


Figure 1.2: Time scales of the main modes of variability at different time scales (top; figure adapted from Merryfield *et al.*, 2020) and their approximate location (bottom).

it is related to impacts on local temperature, precipitation and extreme events in Europe, Asia and North America.

- Southern Annular Mode (SAM; Kidson, 1988), also known as the Antarctic Oscillation: It is characterised mostly through the pressure differences between mid and high latitudes, and is associated with a longitudinal oscillation of the westerly winds over the mid- and high-latitudes of the Southern Hemisphere, influencing the temperature and precipitation over southern Australia.

- El Niño - Southern Oscillation (ENSO; McPhaden *et al.*, 2006): It is characterised by variations in the sea surface temperatures (SST) over the tropical Pacific Ocean, and is related to significant impacts on climate conditions and extremes over large areas of the globe at seasonal time scales. It is considered as the globally dominating mode of variability on inter-annual time scales.

- Quasi-Biennial Oscillation (QBO; Kim & Chun, 2015): downward-propagating oscillation with an approximate duration of 28-month of easterly and westerly zonal jets in the tropical stratosphere, driven by upward equatorial waves from the troposphere. When the QBO is easterly, sudden stratospheric warmings and colder-than-normal winters over Northern Europe are more likely to occur.

- Indian Ocean Dipole (IOD; Ashok *et al.*, 2001), also known as the Indian Niño: It is characterised by changes in SST over the Indian Ocean, and is related to climate extremes over Australia and parts of Asia.

- Interdecadal Pacific Oscillation (IPO; Henley *et al.*, 2015): It is associated with a tripole pattern of SST over the Pacific Ocean, and has regional impacts over regions such as North America, East Asia, South America and Australia.

- Pacific Decadal Oscillation (PDO; Mantua & Hare, 2002): It is characterised by oscillations of the SST over the North Pacific Ocean, and it could be described as a decadal-scale ENSO pattern (but also including extra-tropical SST variations, whereas ENSO is tropical). It is related to impacts over North America and Asia, although it has also been related to impacts in the Southern Hemisphere.

- South Pacific Ocean Dipole (SPOD; Saurral *et al.*, 2020): It is modulated by ENSO and the IPO, and is related to anomalies of temperature and precipitation over parts of the Southern Hemisphere.

- Atlantic Multidecadal Variability (AMV; Doblas-Reyes *et al.*, 2013; Trenberth & Shea, 2006): It is characterised by variations of the SST over the North Atlantic Ocean. It is related to impacts over Europe, Africa and North and Central America.

In addition to variations caused by the modes of internal variability, the climate system also responds to external forcings, generating complex feedbacks and mechanisms (Brovkin *et al.*, 2004). For instance, human-induced greenhouse gas emissions, such as carbon dioxide, have a significant impact on the system, leading to an increase in temperatures, sea-level rise, changes in precipitation patterns, and alteration of natural ecosystems (Jain, 1993); volcanic eruptions, which release volcanic gases and aerosols into the atmosphere, producing a cooling effect during a few years due to the reflecting of sunlight (Hermanson *et al.*, 2020); land use changes (Ward *et al.*, 2014); and variations in the solar radiation lead to small changes in the global temperature (Zanchettin, 2017).

1.2. Climate predictions and projections

Given the variability of the climate system and its potential consequences, predictions of its evolution are crucial for applications in all time scales in order to anticipate and minimise the risk associated with weather and climate events that can cause devastating effects on natural ecosystems and human activities.

The first weather forecast was created by combining ship observations and the meteorologist's intuition, and was issued in the British newspaper *The Times* by Robert FitzRoy in 1861 (Talman, 1927). These forecasts aimed to help sailors avoid dangerous conditions at sea and were distributed in newspapers and public places. Over time, the knowledge about the weather and its evolution has improved, increasing the trustworthiness of weather predictions and boosting their applications by population and organisations on, for example, agriculture, transportation and outdoor activities. Nowadays, weather predictions are performed with numerical climate models, software tools based on mathematical representations that simulate the evolution of the atmosphere by solving the laws of physics (Bauer *et al.*, 2015). The first numerical weather forecast was produced in 1950 using a computer to run a dynamical model. Nowadays, these weather forecasts, which are started from the current atmospheric conditions, provide predictions of the atmospheric evolution from the next hours up to several days.

For longer forecast periods, subseasonal and seasonal climate predictions are typically developed by coupling ocean and atmosphere numerical models to account for their interaction when predicting the climate evolution from the current state. Besides, other components, such as a sea ice model, can also be coupled to incorporate more information when producing the predictions (Vitart *et al.*, 2017). Subseasonal predictions cover a forecast period from one week to a few months and were started in the 1980s. However, the most significant advances were made in the 2000s when more computational resources were available. On the other hand, seasonal predictions provide forecasts for the next months up to a year and were performed from the 1980s when coupled ocean-atmosphere models were developed. Subseasonal and seasonal predictions benefit from the predictability given by different low-frequency modes of variability, such as the MJO and ENSO, among others (Vitart & Robertson, 2018).

On multi-decadal to centennial time scales, long-term climate projections provide information on how the climate is expected to change and its potential impacts in the long-term future under different Shared Socioeconomic Pathways (SSPs; IPCC, 2023a; O'Neill *et al.*, 2020; Riahi *et al.*, 2017), also known as socio-economic scenarios. These scenarios are based on different radiative forcings that may be reached under assumptions of future societal and economic developments, different levels of greenhouse gas emissions, aerosol concentrations, land use and other factors that affect the radiative balance of the climate system (IPCC, 2023a).

Weather, subseasonal and seasonal climate predictions are considered an initial value problem, as most of the predictability is given by the current state of the climate system. In contrast, the predictability for climate projections is provided by external forcings, and thus, they are considered a forced boundary condition problem.

1.3. Decadal climate predictions

Decadal predictions have recently been developed as a source of climate information at annual to decadal time scales, filling the gap between seasonal predictions and climate projections (Doblas-Reyes *et al.*, 2013; Kushnir *et al.*, 2019). The first decadal prediction was made in 2007 by the Met Office Hadley Centre (Smith *et al.*, 2007) by incorporating information about the current climate state into a climate model previously used to produce long-term climate change projections. Thus, decadal climate predictions incorporate information on both the internal climate variability and external forcings (Kirtman *et al.*, 2013; Meehl *et al.*, 2009), benefiting from both sources

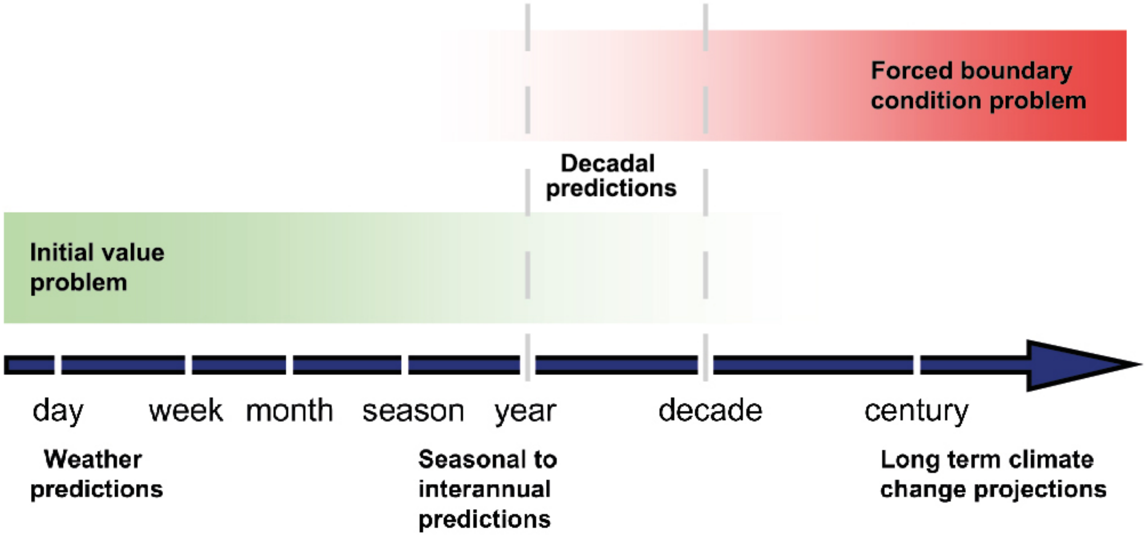

Forced boundary condition problem
Decadal predictions
Initial value problem
day
week
month
season
year
decade
century
Weather predictions
Seasonal to interannual predictions
Long term climate change projections

Figure 1.3: Weather and climate prediction time scales, and the impact by the initial value and forced boundary condition problems on their predictability. Figure taken from Kirtman *et al.* (2013).

of predictability, and are thus considered as a combination of initial value and forced boundary condition problems (Figure 1.3).

Although the term "decadal" may seem to only refer to a whole decade, decadal predictions are also used to provide climate information at shorter time scales. For instance, they can be used to produce annual, multi-annual and multi-seasonal averages (Dunstone *et al.*, 2020; Sospedra-Alfonso & Boer, 2020). At these time scales, high-quality and reliable climate information is crucial to prepare adaptation and mitigation policies to minimise the impact of climate variability and change. The anticipation of climate variations is also beneficial to plan and help decision-making about investments, infrastructure and disaster preparedness, as well as estimating the risk and opportunities associated with such variations of the climate conditions in several climate-vulnerable and socio-economic sectors (BrunoSoares *et al.*, 2018; Merryfield *et al.*, 2020). Besides, decadal predictions can also help to advance climate science by further understanding how the climate system works and the interactions between its different components.

Decadal predictions are created with the same climate models as the long-term climate projections. The major difference between these two types of simulations is that, while climate projections only incorporate information about the external forcings, decadal predictions incorporate information on both external forcings and internal climate variability (Kirtman *et al.*, 2013; Meehl *et al.*, 2009). To do so, the best estimate of the current state of the climate system is incorporated into the climate models to align the phase of the model simulations with that observed (Volpi *et al.*, 2021). This procedure is known as model initialisation (Polkova *et al.*, 2019) and is the main difference between decadal predictions and climate projections (Smith *et al.*, 2019). The two main initialisation methods are the full-field and anomaly initialisation methods (Hazeleger *et al.*, 2013; Polkova *et al.*, 2019; Smith *et al.*, 2013). After the model initialisation date, known as start date and typically selected towards the end of each calendar year, the evolution of the system is computed by integrating the dynamical and physical equations forward in time predicting the future climate evolution up to ten years. As the prediction evolves in time, the information incorporated through the initial conditions is lost, and decadal predictions become more similar to uninitialised climate projections.

The information on the current climate state used for model initialisation, known as initial conditions, is created from observation-based products, which are not perfect (Zumwald *et al.*, 2020). For instance, human and instrumental errors during the data measurement, biases during the data processing and homogenisation to combine different sources of observations, and limited temporal and spatial coverage decrease the quality of the observational data. Therefore, imperfections in the observations are propagated to initial conditions and are then propagated to the predictions due to the chaotic nature of the climate system (Lorenz, 1963). In order to account for this observational uncertainty, climate models are run several times following the ensemble approach, in which each run is referred to as an ensemble member.

There are several ensemble generation methods used to generate spread (i.e. an estimation of the forecast uncertainty) to sample the uncertainty of the initial state of the predictions. For instance, the burst method consists of initialising the forecast system on the same calendar day with small perturbations in the initial conditions. A simpler approach is the lagged method, which consists of initialising the forecast system on slightly different dates (Chen *et al.*, 2013) and has been shown to generate enough spread in the predictions (Polkova *et al.*, 2019). Both approaches allow obtaining the uncertainty of the predictions associated with the initial state. For the first lead times, the predictions usually point to the same trajectory of climate evolution, and the spread between the ensemble members increases for longer lead times, thus increasing the uncertainty of the predictions (Hawkins & Sutton, 2009). This increase in time of the ensemble spread can be seen in the illustration presented in Figure 1.4, with examples of initialised decadal predictions and their spread represented in green colours. In addition, it shows the different behaviour of initialised decadal and uninitialised historical simulations (i.e. climate projections of the past, which use prescribed forcings), with a narrowed uncertainty shown by decadal predictions due to the information provided by initial conditions during the model initialisation, which constrains the spread particularly for the first lead times. The coupled climate model and the complete procedure to create the predictions (including the used forcings, parameterisations, initial conditions and initialisation procedure, among others) constitute the forecast system.

Despite all the recent efforts by the community, there are still many open issues associated with decadal predictions (Merryfield *et al.*, 2020). For example, increasing the model spatial resolution to solve physical processes instead of parameterisations can reduce predictions' biases and improve the forecast quality (Jia *et al.*, 2015; Müller *et al.*, 2018; Schuster *et al.*, 2019). Another of the main challenges in decadal prediction research is to understand dynamical phenomena further and improve their representation and teleconnections in climate models (Cassou *et al.*, 2018). A significant issue of decadal predictions is the initialisation shock that can rapidly occur after model initialisation due to inconsistencies between the initial conditions and the model itself, causing an adverse effect of model initialisation (Bilbao *et al.*, 2021; Kröger *et al.*, 2018; Pohlmann *et al.*, 2017; Volpi *et al.*, 2021). Improving data assimilation methods to incorporate the initial conditions in the climate models may enhance the quality of the predictions and minimise the effects of initialisation shocks. For instance, strong data assimilation methods, which refer to using a common data assimilation for all the coupled model components, may reduce imbalances between the different components after initialisation (Pohlmann *et al.*, 2023).

In addition to the challenges mentioned above, the signal-to-noise issue is currently regarded as one of the major problems related to climate prediction. This issue refers to the low signal-to-noise ratio (i.e., the magnitude of the predictable signal relative to unpredictable noise) in climate predictions (Eade *et al.*, 2014), and it is related

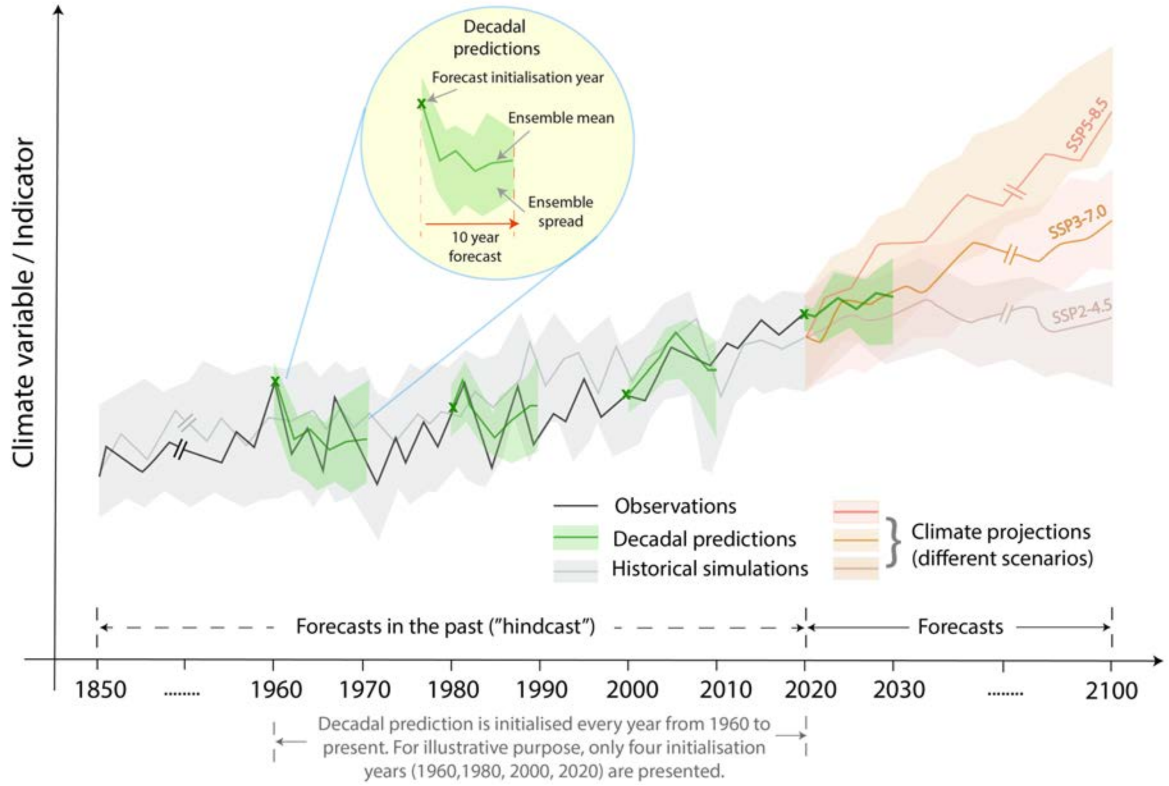


Figure 1.4: Illustration of time series for observations (black line), historical simulations (grey line and shading), climate projections (different intensities of orange lines and shadings) and decadal predictions (green line and shading). Figure taken from Solaraju-Murali *et al.* (2022)

to the signal-to-noise paradox (Scaife & Smith, 2018). The signal-to-noise paradox refers to the fact that models are able to better predict the real world than themselves (as the ensemble mean is typically higher correlated with observations than with the members of the ensemble), and has been shown to occur, for instance, for predictions of the high-latitude blocking and the NAO (Athanasiadis *et al.*, 2020; Scaife & Smith, 2018). Besides, (Smith *et al.*, 2020) showed that the predictable signal in climate models is smaller than in reality, implying that real-world climate evolution is more predictable than the climate model output suggests. Also, the signal-to-noise issue can be minimised using the ensemble mean of very large ensembles to extract the predictable signal from the chaotic noise (Athanasiadis *et al.*, 2020; Donat *et al.*, 2023;

Smith *et al.*, 2020).

1.4. Forecast quality assessment

Before using any climate information derived from the predictions, it is essential to assess the forecast quality (also referred to as skill; defined as the degree of correspondence between the predicted and observed values) to estimate how trustable the climate information is and to assess the potential value of climate predictions (Goddard *et al.*, 2012, 2013). Thus, the prediction skill is assessed by measuring the level of agreement between the predictions and observations. To do so, the same forecast systems used to predict the future variations of the climate system are used to predict such variations over the past decades, when observations are available so that they can be used as reference for comparison. These climate forecasts of past climate conditions are called climate hindcasts or climate re-forecasts.

There are a number of skill metrics that can be used to evaluate the quality of climate predictions. However, the specific metric should be chosen depending on the particular forecast product that is being evaluated. Alternatively, in a pure research context, different metrics should be used to cover all aspects that may be relevant for users (Goddard *et al.*, 2012). A forecast product can be deterministic or probabilistic. Deterministic forecast products are based on the ensemble mean or median. In contrast, probabilistic forecast products are based on probabilities estimated as the fraction of ensemble members that fall into each discrete probabilistic category (e.g. equiprobable tercile or quintile categories) or on the full probability distribution function. The forecast quality can also be provided in terms of skill scores, which measure the quality of a forecast in comparison to a reference forecast and allow easily identifying whether a forecast outperforms a reference forecast (Goddard *et al.*, 2013). For instance, skill scores can be used to compare different versions of the same model, different models, initialised and uninitialised experiments, and model output against classical forecasts such as those based on climatology or persistence (Murphy, 1992).

In addition to the forecast quality, assessing the forecast reliability is essential, as it is crucial for decision-making procedures with climate information. A prediction is reliable if the ensemble spread is representative of the forecast uncertainty (Verfaillie *et al.*, 2021). For example, the deterministic forecast is reliable if the spread matches the error, meaning the predictions fall into the forecast uncertainty. On the other hand, a probabilistic forecast is reliable if there is an agreement between forecast probability and mean observed frequency (Weisheimer & Palmer, 2014). In a practical example, a probabilistic forecast is reliable if it rained 20% of the times that there was a probability of 20% of rainy conditions.

Given the observational uncertainty, the skill estimates differ when measured against observation-based datasets. Also, some reanalysis may have been created using models included as a component in coupled climate models, and they may therefore have common biases and overestimate the actual skill of such models. Thus, using more than one observational dataset is recommended when performing the forecast quality assessment (Goddard *et al.*, 2013; Jolliffe & Stephenson, 2012). Using several observational reference datasets allows for comparing the skill estimates to check the robustness of the results. Other options include using the mean or the median of several reference datasets to make a robust forecast evaluation or using a reference dataset that provides uncertainty estimates based on the standard error (e.g. the Global Precipitation Climatology Project dataset, GPCP; Adler *et al.*, 2003) or based on an ensemble (e.g. the 5th generation ECMWF Atmospheric Reanalysis, ERA5; Hersbach *et al.*, 2020).

1.5. Post-processing techniques

In addition to their use in forecast quality assessment, climate hindcasts are also needed to post-process the raw model output to partially correct systematic errors that models have, known as model biases, and to make the statistical properties of the forecasts more similar to those observed. These errors are caused, for instance, by imperfections in the model, inconsistencies between the model and the observation products used to initialise the predictions which lead to initialisation shocks, approximations during the mathematical representation and limited knowledge about some climate system's processes. The simplest bias adjustment consists of removing the model climatology and adding the observed climatology, correcting the mean bias in the predictions (Gangstø *et al.*, 2013). Besides, biases in the variance can also be corrected by multiplying by the ratio of the observed and predicted variances (Torralba *et al.*, 2017). In addition to biases in the mean and variance, other statistical properties, such as the spread, can be corrected by applying more advanced calibration methods with the goal of improving the reliability of the predictions (Doblas-Reyes *et al.*, 2005; Eade *et al.*, 2014; Manzanas *et al.*, 2019; Marcos *et al.*, 2018; Pasternack *et al.*, 2018; Pérez-Zanón *et al.*, 2021; Schaeybroeck & Vannitsem, 2011, 2015; Smith *et al.*, 2020; Tippett & Barnston, 2008; Torralba *et al.*, 2017; Weigel *et al.*, 2009; Zhao *et al.*, 2017). Every calibration method should be applied in cross-validation mode, which consists of not using observed information for the time step that is being calibrated, which is essential when bias-adjusting or calibrating hindcasts not to produce overfitting and thus overestimate the actual skill of future predictions (Doblas-Reyes *et al.*, 2005). In contrast, cross-validations can also underestimate the true skill, particularly when correcting a small sample size (Gangstø *et al.*, 2013; Smith *et al.*, 2013).

Running climate models to create the predictions is highly computationally expensive, which limits the spatial resolution at which they are run. Therefore, such a spatial resolution is often insufficient to address the users' needs, as the grid may cover a too large area to represent the climate variations over a specific location. Statistical downscaling methods are post-processing methods to create more localised information by increasing the forecasts' spatial resolution to improve the climate information by better understanding the potential impacts at a regional scale (Benestad *et al.*, 2019; Paxian *et al.*, 2022).

There is also uncertainty associated with the forecast system as none of them are perfect. The multi-model ensemble method combines predictions from several forecast systems into a single ensemble. The multi-model ensemble has been shown to outperform the predictions from a single forecast system of the same ensemble size (DelSole *et al.*, 2014) and can outperform the best single forecast system that the multi-model ensemble contains (Bellucci *et al.*, 2014; Hagedorn *et al.*, 2005; Hemri *et al.*, 2020). The higher skill of the multi-model ensemble is due to the increased ensemble size, the error cancellation between models, and the signal each forecast system adds to the multi-model ensemble (Hagedorn *et al.*, 2005). However, the multi-model does not necessarily show higher skill than individual forecast systems for all cases (Mishra *et al.*, 2018).

1.6. Climate services

Climate services provision implies collecting, analysing and disseminating climate information so that it allows to support decision-making and planning strategies to anticipate, adapt, manage disaster risk and minimise the impacts of climate variability and change on communities and climate-vulnerable sectors (Merryfield *et al.*, 2020). For instance, these sectors include agriculture (Solaraju-Murali *et al.*, 2021), renewable energy (BrunoSoares *et al.*, 2018), water management (Paxian *et al.*, 2019), health (Frumkin *et al.*, 2011), fisheries (Tommasi *et al.*, 2017), insurance (Caron *et al.*, 2018), retail (Chiang & Ling, 2017), and urban planning (Masson *et al.*, 2020). Decadal climate predictions have only been used for pure scientific research until very recently (Solaraju-Murali *et al.*, 2022). However, in 2020, the Copernicus Climate Change Service (C3S) promoted a collaboration between four European institutions (Deutscher Wetterdienst, DWD; Barcelona Supercomputing Center, BSC; Centro Euro-Mediterraneo per i Cambiamenti Climatici, CMCC; and United Kingdom's Met Office; UKMO) to create and deliver prototypes of climate services based on decadal predictions for the agriculture, energy, infrastructure and insurance sectors (https://climate.copernicus.eu/sectoral-applications-decadal-predictions).

In order to deliver a climate service, the climate information produced with the forecast systems needs to be translated into reliable, tailored and actionable information that enable stakeholders and decision-makers to take action. To do so, a co-production process involving both scientists and decision-makers is crucial to build trust and understand the specific user's needs and requirements that the final product must meet (Bojovic *et al.*, 2021; Goddard *et al.*, 2012). This knowledge exchange serves as the basis for developing co-designed new climate information, and its added value and potential applicability can be addressed through, for instance, support decision tools or case studies (Solaraju-Murali *et al.*, 2022) to demonstrate the usefulness of the provided climate information.

After identifying, selecting and engaging the users that will be provided with the climate service (Baulenas *et al.*, 2023), there are six main steps in the co-production process to ensure tailored and usable climate information (Solaraju-Murali, 2023). The first step is to understand the user's requirements and needs, in which users and climate services providers exchange their knowledge and understand the decisions the potentially provided climate information would support. For example, users can state which variables or indicators, forecast periods, regions and final products are required to be incorporated into their decision-making process (BrunoSoares *et al.*, 2018). The second step is identifying and obtaining the relevant climate data for the climate service. This data collection is constrained by the availability and timeliness of the climate predictions produced with the forecast systems, for example, on the Earth System Grid Federation (ESGF; https://esgf-node.llnl.gov), which serves as the database for the decadal prediction data produced within the framework of the Decadal Climate Prediction Project (DCPP; Boer *et al.*, 2016) component of the Coupled Model Intercomparison Project Phase 6 (CMIP6; Eyring *et al.*, 2016).

After collecting the necessary climate data, the third step implies applying post-processing methods, such as multi-model combination, bias adjustment, calibration and downscaling, aiming to improve the quality and reliability of the raw model output (see Section 1.5). The fourth step consists of evaluating the forecast quality and reliability of these post-processed predictions to verify whether they show quality and demonstrate value to inform users and decision-makers (see Section 1.4). These two steps should be carried out in an iterative process since different post-processing techniques may be used and compared to select the method that provides the highest quality information for each specific case.

The fifth step involves creating tailored climate information by producing climate indices or indicators for the specific user's needs and decisions to be taken. These indices and indicators can also be understood as another post-processing method, as they are computed with one or more of the essential climate variables (e.g. monthly means of temperature and precipitation) from the model output. Examples of these tailored indices and indicators are the Standardised Precipitation Evapotranspiration Index (SPEI; Vicente-Serrano *et al.*, 2010) and other indices such as those recommended by the Expert Team on Climate Change Detection and Indices (ETCCDI; Zhang *et al.*, 2011) but adapting them using user-defined thresholds for them to be as tailored as possible to the user's needs. Finally, the sixth and last step is developing a climate service product with such tailored information. Such a product must be understandable, trusted and usable by the user to be incorporated into their decision process (Bojovic *et al.*, 2021). It can be presented differently, such as maps or time series included in websites, apps, text documents or a combination. These steps of the co-production process can be repeated as many times as needed in order to keep including the user's feedback and finally deliver the most useful climate information during climate service provision.

1.7. Framework within the Earth Sciences Department at the BSC-CNS

This Ph.D. thesis has been developed within the Climate Variability and Change (CVC) and the Earth System Services (ESS) groups of the Earth Sciences (ES) department at the Barcelona Supercomputing Center - Centro Nacional de Supercomputación (BSC-CNS).

The CVC group, co-led by Markus G. Donat and Pablo Ortega, focuses on further understanding climate variability and its sources of predictability to increase the model accuracy in representing the physical and biogeochemical processes in climate models. The group works on a wide range of research activities related to, for example, data assimilation, model initialisation, ensemble generation and timescales-merging methods to improve the forecast quality of the EC-Earth forecast system. The BSC is one of the members of the EC-Earth Consortium (https://ec-earth.org/consortium/), which collaborates on the development of the model. Besides, the BSC is a Global Producing Centres, designed by the World Meteorological Organization (WMO), that provides the WMO Lead Centre for Annual-to-Decadal Climate Prediction (https://hadleyserver.metoffice.gov.uk/wmolc/) with timely decadal predictions, which also contribute to CMIP6/DCPP-A.

The ESS group, led by Albert Soret, focuses on transferring scientific knowledge and advancing sustainable development in different sectors (e.g. renewable energy, agriculture, health, urban development, insurance and water management), combining the knowledge from scientific and social researchers. The group works on the interpretation, communication and applicability of scientific research, and improving the forecast quality by applying post-processing techniques. The final aim is to provide climate services at different time scales to benefit nature and society by conducting and improving user interaction to create the most tailored services.

This Ph.D. thesis has benefitted from its development within both groups, one focused on fundamental climate research and the other on application-oriented research. The production and application of decadal predictions are supported by the teamwork within the department, starting from the production of the predictions (documented in Bilbao *et al.*, 2021) to the delivery of climate information to stakeholders (e.g. the climate services provision for the wheat sector, documented in Solaraju-Murali *et al.*, 2021). Previous Ph.D. theses have also supported the research conducted within this Ph.D., particularly the ones carried out by Verórica Torralba and Balakrishnan Solaraju-Murali. Torralba (2019) developed a Ph.D. on the applicability of seasonal predictions for the wind energy sector. Her work included the analysis of Euro-Atlantic weather regimes on seasonal predictions, and was the basis for one of the research articles of this Ph.D. thesis. Solaraju-Murali (2023) developed his Ph.D. thesis on the applicability of decadal predictions to support decision-making in the agriculture sector, and has been the basis for the forecast product development within several projects and contracts (presented in Chapter 6).

The research presented in this thesis has also been fostered by collaborations in several Spanish (CLINSA) and European (C3S_34c, EUCP, FOCUS-Africa, ASPECT) projects, as well as a contract with a private company (Decathlon). The specific tasks developed within each project and contract are stated in Appendix A.7. Having the opportunity to interact and collaborate with climate researchers from BSC-CNS and other international institutions has favoured the development and learning process of this thesis, also giving the opportunity to co-authoring some research articles listed in Appendix A.2.

Chapter 2

Objectives and structure

The overall objective of this Ph.D. thesis is to evaluate the quality of decadal climate predictions and investigate their potential usefulness for climate services at annual to decadal time scales. The forecast quality assessment allows for finding windows of opportunity in which climate predictions show skill or added value in comparison to reference forecasts (e.g. the climatological forecast) to be used for climate services provision in several climate-vulnerable sectors such as agriculture, energy, water management and health at local, regional and global levels. Besides, post-processing techniques such as calibration and downscaling methods and multi-model approaches are applied to the raw forecasts with the aim of improving the quality, reliability and thus usability of the climate information for it to be applicable for policy and decision-making processes.

In particular, predictions of mean variables (e.g. near-surface air temperature and precipitation), modes of climate variability (e.g. the AMV and PDO indices), extreme climate indices (e.g. the most extreme days in terms of cold and hot temperatures and heavy rainfall), drought-related indices (e.g. SPEI index) and Euro-Atlantic weather regimes (e.g. NAO and Blocking regimes) have been evaluated to seek for opportunities of climate services provision in several sectors, and as well as to detect model deficiencies for further improvements in the next generation of forecast systems.

Furthermore, the impact of model initialisation towards the observed climate state is also assessed by comparing the initialised decadal hindcasts and uninitialised historical forcing simulations contributing to CMIP6. This comparison allows for selecting the highest-quality climate information for each specific case, as decadal predictions and climate projections may show a different quality depending on the particular case being analysed. Also, the estimation of the impact of model initialisation is needed to assess whether yearly producing decadal predictions is worth the effort with regard to the very high computational cost involved.

This Ph.D. thesis consists of seven chapters and four appendices. Chapter 1 introduces the current knowledge on decadal climate predictions, their quality assessment and post-processing approaches with a particular focus on their applicability to climate services. Chapter 2 presents the main objectives of the thesis and how it is structured. Chapters 3, 4 and 5 are shown in the form of a compendium of research articles published in some of the leading peer-reviewed journals in the field (i.e. Journal of Climate, Journal of Geophysical Research-Atmospheres, and Environmental Research Letters). Chapter 6 discusses the research outcomes and gives an overview of how the research and software developed during the research conducted within this thesis has been applied within Spanish and European projects. Chapter 7 summarises the main conclusions and presents future research. Additionally, Appendix A provides a list of the authored and co-authored papers, attendance to conferences and workshops, contribution to projects and software development, and information on the research stay conducted at the International Research Institute for Climate and Society (IRI; https://iri.columbia.edu/) of Columbia University during the Ph.D.; and Appendices B, C and D include the supplementary materials of the research articles shown in Chapters 3, 4 and 5, respectively.

Chapter 3

Multi-Model forecast quality assessment of CMIP6 decadal predictions

This chapter has been published as peer-reviewed article as:

Delgado-Torres, C., Donat, M. G., Gonzalez-Reviriego, N., Caron, L., Athanasiadis, P. J., Bretonnière, P., Dunstone, N. J., Ho, A., Nicoli, D., Pankatz, K., Paxian, A., Pérez-Zanón, N., Cabré, M. S., Solaraju-Murali, B., Soret, A., and Doblas-Reyes, F. J. (2022). Multi-Model forecast quality assessment of CMIP6 decadal predictions. Journal of Climate, 35(13), 4363-4382. https://doi.org/10.1175/JCLI-D-21-0811.1

The supplementary material can be found in Appendix B.

3.1. Main objectives

- Evaluate the forecast quality of the decadal predictions contributing to CMIP6/DCPP in predicting near-surface air temperature, precipitation, the AMV index and the GSAT anomalies.

- Assess whether the approach used to build a multi-model ensemble has a significant impact on the prediction skill.

- Compare the multi-model ensemble against the individual forecast systems to assess the benefit and drawbacks of combining predictions from different forecast systems.
- Estimate the impact of model initialisation by comparing the skill of the DCPP and historical forcing simulations (HIST) multi-model ensembles.
- Estimate how much skill is lost for not having all the predictions available in real-time by comparing a research multi-model ensemble (13 forecast systems) against an operational multi-model ensemble (4 forecast systems).

3.2. Main outcomes

- The DCPP multi-model ensemble is skilful in predicting near-surface air temperature for the forthcoming five years over most of the global domain.
- The DCPP skill in predicting precipitation is lower in comparison to temperature and significant skill is only found over regions of Central Africa, Europe, and Asia.
- The DCPP multi-model ensemble skillfully predicts both the AMV index and the GSAT anomalies.
- The four approaches used to build a large multi-model ensemble show a similar skill for all the considered variables and indices.
- The best forecast system generally provides the highest skill for a particular location, variable and forecast period, indicating that it is the best option for a particular climate service.
- The multi-model ensemble shows higher skill than, at least, the 50% of the individual forecast systems, thus being a reasonable choice for operational forecast generation.
- There is an added value from model initialisation for predictions of temperature and precipitation over some ocean and land regions. Also, there is added value for the AMV index and GSAT anomalies.

- The research multi-model ensemble shows generally higher quality than the operational multi-model ensemble, although the differences are not always statistically significant. This suggests that more real-time predictions would be beneficial and allow selecting the best forecast system or multi-model ensemble for each specific region, variable and forecast period.

Multi-Model Forecast Quality Assessment of CMIP6 Decadal Predictions

Carlos Delgado-Torres,[a] Markus G. Donat,[a,b] Nube Gonzalez-Reviriego,[a] Louis-Philippe Caron,[a,c] Panos J. Athanasiadis,[d] Pierre-Antoine Bretonnière,[a] Nick J. Dunstone,[e] An-Chi Ho,[a] Dario Nicoli,[d] Klaus Pankatz,[f] Andreas Paxian,[f] Núria Pérez-Zanón,[a] Margarida Samsó Cabré,[a] Balakrishnan Solaraju-Murali,[a] Albert Soret,[a] and Francisco J. Doblas-Reyes[a,b]

[a] *Barcelona Supercomputing Cente, Barcelona, Spain*
[b] *Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain*
[c] *Ouranos, Montreal, Quebec, Canada*
[d] *Centro Euro-Mediterraneo sui Cambiamenti Climatici, Bologna, Italy*
[e] *Met Office Hadley Centre, Exeter, United Kingdom*
[f] *Business Area of Climate and Environment, Deutscher Wetterdienst, Offenbach (Main), Germany*

(Manuscript received 13 October 2021, in final form 3 March 2022)

ABSTRACT: Decadal climate predictions are a relatively new source of climate information for interannual to decadal time scales, which is of increasing interest for users. Forecast quality assessment is essential to identify windows of opportunity (e.g., variables, regions, and forecast periods) with skill that can be used to develop climate services to inform users in several sectors and define benchmarks for improvements in forecast systems. This work evaluates the quality of multi-model forecasts of near-surface air temperature, precipitation, Atlantic multidecadal variability index (AMV), and global near-surface air temperature (GSAT) anomalies generated from all the available retrospective decadal predictions contributing to phase 6 of the Coupled Model Intercomparison Project (CMIP6). The predictions generally show high skill in predicting temperature, AMV, and GSAT, while the skill is more limited for precipitation. Different approaches for generating a multi-model forecast are compared, finding small differences between them. The multi-model ensemble is also compared to the individual forecast systems. The best system usually provides the highest skill. However, the multi-model ensemble is a reasonable choice for not having to select the best system for each particular variable, forecast period, and region. Furthermore, the decadal predictions are compared to the historical simulations to estimate the impact of initialization. An added value is found for several ocean and land regions for temperature, AMV, and GSAT, while it is more reduced for precipitation. Moreover, the full ensemble is compared to a subensemble to measure the impact of the ensemble size. Finally, the implications of these results in a climate services context, which requires predictions issued in near–real time, are discussed.

KEYWORDS: Climate prediction; Ensembles; Forecast verification/skill; Hindcasts; Probability forecasts/models/distribution; Decadal variability; Climate services

1. Introduction

Decadal climate prediction aims to predict the evolution of the climate system from 1 to 10 years ahead, filling the gap between seasonal predictions and climate projections. Although the word "decadal" may seem to only refer to a whole decade, these predictions also include shorter time scales, as the forecasts are typically issued for annual, multiannual, and multiseasonal averages (Dunstone et al. 2020; Sospedra-Alfonso and Boer 2020), which strongly depends on the user's needs. The external forcings (natural and anthropogenic) and the internal climate variability (i.e., the slow variations of the climate system) provide predictability on these time scales (Doblas-Reyes et al. 2013; Goddard et al. 2013; Smith et al. 2019). However, due to chaotic characteristics of the climate system, it is not possible to predict its exact evolution. Thus, decadal forecasting provides large ensembles of predictions (needed for sampling observational uncertainty in the initial conditions and increasing the signal-to-noise ratio; Smith et al. 2020), which, besides predicting the average anomalies based on the ensemble mean, are also used to obtain probabilistic information about the likelihood for certain event types to occur.

Murphy (1993) presented three ways to evaluate a forecast: forecast quality (defined as the degree of correspondence between the simulated and observed conditions), consistency (which is based on the forecaster's judgments and knowledge), and value (the usefulness that the forecasts can provide to increase social, economic, environmental, or other benefits when used by users and decision-makers). Knowing the quality of a prediction is crucial for providing useful forecast products for specific sectors such as the agriculture (Solaraju-Murali et al. 2021), energy (Bruno Soares et al. 2018), water management (Paxian et al. 2019), health (Frumkin et al. 2008), marine fisheries (Tommasi et al. 2017), and insurance sectors (Caron et al. 2018), as well as improving the current forecast systems (Vera et al. 2010; Goddard et al. 2013; Fricker et al. 2013).

The forecast quality assessment needs to be carried out over a long enough period in the past (when observations are

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JCLI-D-21-0811.s1.

Corresponding author: Carlos Delgado-Torres, carlos.delgado@bsc.es

DOI: 10.1175/JCLI-D-21-0811.1

available to compare against) to achieve robust results that can be used as an estimate of how well the forecast system may perform in simulating future climatic anomalies. Thus, retrospective decadal forecasts (also termed *hindcasts* or *reforecasts*) are performed with the same forecast systems used to predict future climate variations. For this, the forecast systems are utilized to simulate the evolution of the climate system from a set of initial conditions based on observations, which is referred to as forecast system initialization (Hazeleger et al. 2013; Smith et al. 2013; Polkova et al. 2019), and incorporate information about the external forcings. These decadal forecasts stand in contrast to retrospective climate projections (known as *historical simulations*), which incorporate the same external forcings as in the decadal forecasts, but which do not align the internal variability of the forecast system with that of the real world through an initialization procedure. The comparison between decadal hindcasts and the historical simulations performed using the same climate model and ensemble size provides an estimate of the impact of the forecast system initialization on the quality of the predictions (Doblas-Reyes et al. 2013; Smith et al. 2019).

Errors in the initial conditions (generated from observation-based data) and errors in the model itself (e.g., imperfect representation of the climate system and approximation in mathematical methods to solve the physics equations) can affect the quality of the forecasts (Slingo and Palmer 2011). The ensemble approach, which consists of performing several simulations from slightly different initial conditions or by perturbing parameters of the forecast system, is commonly used to address these limitations. Each one of these simulations is called an ensemble member and provides a possible evolution of the climate system. The ensemble approach also allows obtaining an estimate of the forecast system's confidence in its predictions. For instance, the agreement of all members to the same evolution of the climate systems indicates that the forecast system is confident about the climate trajectory it is predicting. By contrast, a wide range of future climate evolutions simulated by the different ensemble members reveals low confidence. It should be noted that even in the case that all the ensemble members were very similar (i.e., the forecast system is confident about its simulations), the forecast may not be of high quality due to internal errors in the forecast system itself, such as the lack of forecast reliability (i.e., the degree of correspondence between the forecast probability and the mean observed frequency for a certain event; Murphy 1993).

All forecast systems suffer from systematic errors. For example, the forecast system's climatology may differ from the observed one. Thus, after initializing the system with initial conditions that differ from the model's preferred state, the predictions will slowly evolve toward the model's preferred state (i.e., its climatology). This is known as the model drift (Boer et al. 2016). Similarly, the variance of the simulated time series can be different from the observed one. Postprocessing techniques allow us to partially correct some of the statistical properties of the raw ensembles of simulations, making them more consistent with the reference dataset. Correcting both the mean and variance of the predictions is a strategy known as simple bias adjustment. More complex postprocessing techniques can partially correct other statistical properties of the predictions, making them more reliable or improving the skill measured by specific metrics (different techniques may improve some metrics but worsen others). This procedure is often referred to as calibration (Doblas-Reyes et al. 2005; Tippett and Barnston 2008; Weigel et al. 2009; van Schaeybroeck and Vannitsem 2011; Eade et al. 2014; van Schaeybroeck and Vannitsem 2015; Torralba et al. 2017; Zhao et al. 2017; Manzanas and Gutiérrez 2018; Marcos et al. 2018; Pasternack et al. 2018; Smith et al. 2020; Pérez-Zanón et al. 2021).

Because of the imperfections of the forecast systems, climate predictions must account for the uncertainties associated with a particular forecast system. The multi-model ensemble method, also known as the multisystem ensemble method, combines predictions derived from several forecast systems. Various studies have proven multisystem ensembles to produce more reliable predictions without compromising their accuracy, outperforming a single forecast system of the same ensemble size (DelSole et al. 2014) and even the best of the single forecast systems included in the multi-model in some cases (Hagedorn et al. 2005; Bellucci et al. 2014; Athanasiadis et al. 2017; Hemri et al. 2020). This improvement through the multi-model approach compared to the performance of a single forecast system is not only due to the increase of the ensemble size and the associated error cancellation but also because of the signal that each forecast system adds to the multi-model (Hagedorn et al. 2005). However, it should be noted that not always a multi-model approach provides better results than a single forecast system (Mishra et al. 2018).

There are many approaches to build a multi-model, but there is no agreement on which is the best (Kirtman and Pirani 2009; Hemri et al. 2020). For example, the simplest way is by averaging the ensemble means from each forecast system to create a deterministic forecast product and by averaging the probability density functions to create a probabilistic product. This approach is known as simple multi-model, in which all the forecast systems have the same weight. Another approach for building a multi-model forecast is by pooling the members of all systems together for computing both the ensemble mean and probabilities for the deterministic and probabilistic products, respectively. In that case, the different forecast systems contribute with a different weight to the multi-model ensemble based on their ensemble size. Both approaches are tested in this study, as in Lledó et al. (2020) for seasonal prediction ensembles. Another possible way to build the multi-model ensemble is by weighting the contribution of the forecast systems based on their performance. However, Mishra et al. (2018) found that an equally weighted multi-model outperformed two different unequally weighted multi-models in seasonal forecasting.

The aim of this paper is to assess the quality of the decadal predictions contributing to the Component A of the Decadal Climate Prediction Project (DCPP-A; Boer et al. 2016), part of phase 6 of the Coupled Model Intercomparison Project (CMIP6; Eyring et al. 2016), assessing the sensitivity of the results to different ways of constructing a multi-model

TABLE 1. Forecast systems contributing to the DCPP-A component of the CMIP6 and their specifications (available simulations at the time of the study). The spatial resolution is shown for the atmospheric grid as degrees latitude × degrees longitude.

Forecast system	Institution	DCPP members	HIST members	Spatial resolution	Initialization month	Reference
BCC-CSM2-MR	BCC	8	3	1.125° × 1.125°	January	Wu et al. (2019)
CanESM5	CCCma	20	40	2.8° × 2.8°	January	Swart et al. (2019)
CESM1-1-CAM5-CMIP5	NCAR	40	40	0.9° × 1.25°	November	Yeager et al. (2018)
CMCC-CM2-SR5	CMCC	10	1	0.9° × 1.25°	November	Cherchi et al. (2019)
EC-Earth3-i1	BSC	10	10	0.7° × 0.7°	November	Bilbao et al. (2021)
EC-Earth3-i2	SMHI/DMI	5	—	0.7° × 0.7°	November	Tian et al. (2021)
HadGEM3-GC3.1-MM	MOHC	10	4	0.55° × 0.83°	November	Sellar et al. (2020)
IPSL-CM6A-LR	IPSL	10	32	1.25° × 2.5°	January	Boucher et al. (2020)
MIROC6	MIROC	10	10	1.4° × 1.4°	November	Tatebe et al. (2019)
MPI-ESM1.2-HR	DWD	10	10	0.9° × 0.9°	November	Müller et al. (2018)
MPI-ESM1.2-LR	DWD	16	10	1.9° × 1.9°	November	Mauritsen et al. (2019)
MRI-ESM2-0	MRI	10	5	1.125° × 1.125°	November	Yukimoto et al. (2019)
NorCPM1	NCC	10	30	1.9° × 2.5°	October	Bethke et al. (2021)

ensemble, and also estimating the benefit of using such a multi-model ensemble instead of individual forecast systems. In addition, the impact that the system initialization and the number of systems used to build a multi-model forecast have on the skill is evaluated.

The paper is structured as follows. Section 2 presents the simulations and the observation-based data used in this study. Section 3 describes the postprocessing techniques, the definition of the indices, and the metrics used to evaluate the quality of the forecasts. Section 4 presents and discusses the skill of different multi-model approaches (section 4a), the comparison between the skill obtained with the multi-model ensemble and the individual forecast systems (section 4b), the impact of the system initialization (section 4c), and the impact that the number of forecast systems have on the quality of the multi-model forecast products (section 4d). Finally, conclusions are drawn in section 5.

2. Data

The available decadal predictions from the forecast systems contributing to the DCPP-A component of the CMIP6 have been used in this study. DCPP-A includes the production and analysis of a large multi-model ensemble of hindcasts to assess and understand the decadal prediction skill. It serves as a basis for improving future decadal predictions and for estimating the quality of potential operational forecast production on annual to decadal time scales, which is included in Component B of DCPP (DCPP-B; Boer et al. 2016). In addition to the decadal predictions, the CMIP6 historical simulations performed with the same forecast systems have been used for comparison and to estimate the impact of initialization. The coupled climate forecast systems are the BCC-CSM2-MR, CanESM5, CESM1-1-CAM5-CMIP5, CMCC-CM2-SR5, EC-Earth3 [using both full-field (i1) and anomaly (i2) initialization], HadGEM3-GC3.1-MM, IPSL-CM6A-LR, MIROC6, MPI-ESM1.2-HR, MPI-ESM1.2-LR, MRI-ESM2-0, and NorCPM1, which make a total of 169 decadal predictions (DCPP) and 195 historical simulations (HIST). The main information about these forecast systems, including their ensemble size (number of members), is shown in Table 1.

In this study, the predictions for the average of years 1 to 5 have been evaluated for the global domain. Since the forecast systems are initialized in different calendar months (see Table 1), the first few forecast months have been discarded for some of the forecast systems in order to restrict the analysis to full calendar years (i.e., from January to December). More precisely, the first two forecast months have been discarded for the systems initialized in November, and the first three forecast months have been discarded for the system initialized in October. Both the decadal predictions and the historical simulations have been evaluated over the 1961–2014 period. The evaluation period finishes in 2014 because the historical simulations are run until that year. Hence, start dates from 1960 to 2009 have been used for the evaluation of the decadal predictions in this study.

Monthly means of three climate variables have been used to evaluate the forecasts: near-surface air temperature (tas), precipitation (pr), and sea surface temperature (tos). To take into account the observational uncertainty, two observation-based datasets per variable have been used as reference for the assessment. The first reference dataset listed in Table 2 for each variable has been used for map figures in the main text (the ones marked with an asterisk), while the maps obtained with the second reference dataset are shown in the online supplemental material. In the case of figures displaying the temporal evolution of the indices, the results obtained using both reference datasets are shown.

3. Methods

The forecast quality assessment has been performed with the anomalies of the considered variables to overcome the systematic errors that arise because the climatologies of the forecast systems and the reference datasets are not the same (a problem directly related to mean model biases). The anomalies are computed with respect to the respective 1981–2010 climatology (reference period used by the WMO Lead Centre for Annual-to-Decadal Climate Prediction; https://hadleyserver.

TABLE 2. Reference datasets used for the forecast quality assessment for each variable. The spatial resolution is shown for the atmospheric grid as degrees latitude × degrees longitude. The datasets marked with an asterisk correspond to those used for the map figures in the main text.

Variable	Dataset	Institution	Type	Spatial resolution	Reference
Near-surface air temperature (tas)	GHCNv4*	NOAA	Gridded observations	5° × 5°	Menne et al. (2018)
Near-surface air temperature (tas)	JRA-55	JMA	Reanalysis	1.25° × 1.25°	Kobayashi et al. (2015)
Precipitation (pr)	GPCC*	DWD	Gridded observations	1° × 1°	Schneider et al. (2018)
Precipitation (pr)	JRA-55	JMA	Reanalysis	1.25° × 1.25°	Kobayashi et al. (2015)
Sea surface temperature (tos)	GISTEMPv4	NOAA/NASA	Gridded observations	2° × 2°	Lenssen et al. (2019)
Sea surface temperature (tos)	HadCRUT4	UEA/MOHC	Gridded observations	5° × 5°	Morice et al. (2012)

metoffice.gov.uk/wmolc/). The results obtained using the whole period for computing the climatology were found to be overall very similar (not shown). For the decadal predictions, monthly lead-year-dependent climatologies have been computed so as to account for the model drift, occurring after initialization, toward the respective climatology (see appendix E of Boer et al. 2016). The use of a lead-time-dependent climatology intrinsically corrects the bias and drift of the forecast system's climatology. The same historical period (1981–2010) has also been used to compute the thresholds between the three equiprobable categories (i.e., below lower tercile, between lower and upper terciles, and above upper tercile) for the probabilistic forecast products. Once the three equiprobable categories are computed, the probabilistic forecast products are created as the percentage of ensemble members that corresponds to each category at each time step. Therefore, the probabilistic products are based on three percentages at each time step, which sum up to 100%.

The calibration method proposed in Doblas-Reyes et al. (2005) has been used to assess the impact of calibration on the quality of the products. This calibration method is based on an adjustment and variance inflation of the predictions. It aims at correcting the bias, the overall forecast variance, and the ensemble spread, increasing the reliability of the forecast. The calibration has been performed in leave-one-out cross-validation mode (i.e., excluding the year in which the prediction is made), as it would be for real-time forecasting (Doblas-Reyes et al. 2005; Torralba et al. 2017; Pasternack et al. 2018).

The forecast systems and reference datasets have different spatial resolutions (see Tables 1 and 2). Hence, it was necessary to interpolate the data to a common spatial grid before performing the quality assessment. To avoid interpolating to higher resolutions, for every specific variable the spatial grid chosen was the coarsest grid among the forecast systems and the reference dataset used for the evaluation. The coarsest grid among the forecast systems is the one from the CanESM5 system (gridpoint resolution: 128 longitude × 64 latitude). Thus, the simulated and observed data have been interpolated to the CanESM5's grid when it is coarser than the reference dataset's grid. Otherwise, the simulated and observed data have been interpolated to the reference dataset's grid. Consequently, a different grid has been selected depending on the reference dataset used for each variable.

While decadal predictions, historical simulations and reanalyses typically provide data for the entire globe, observational datasets have missing data for certain regions and periods. When this occurred, all the simulations were masked during the forecast quality assessment in order to have a consistent coverage with the reference datasets (Cowtan et al. 2015). This masking has been applied over the grid points for which at least one missing monthly value was found during the evaluation period.

The global surface air temperature (GSAT) anomaly has been computed as the area-weighted averaged near-surface air temperature anomalies over the entire globe. In the case of the GSAT computed with the GHCNv4 dataset, which has missing values, the GSAT has been computed as the average of the northern and Southern Hemisphere averages. To give the same importance to both hemispheres (the Northern Hemisphere would dominate otherwise, as it has fewer missing values than the Southern Hemisphere). The Atlantic multidecadal variability (AMV) index (Trenberth and Shea 2006) has been computed as the difference between the area-weighted averaged sea surface temperature anomalies over the North Atlantic region (0°–60°N, 280°–360°E) and those over all longitudes in the 60°S–60°N band, as in Doblas-Reyes et al. (2013). The computation of these indices has been performed in the original grid for both the simulations and reference datasets.

Four different multi-model approaches have been tested in this study in order to assess how the construction of a multi-model product affects the quality of the forecast, and to estimate the advantages and shortcomings of combining simulations from several forecast systems instead of using only one of them. The various multi-model ensembles were built as follows:

- Multi-model-1 was built by averaging the individual ensemble means of the different forecast systems when creating deterministic products and by averaging the probabilities computed from the individual forecast systems in the case of probabilistic products. This approach is known as the simple multi-model ensemble, and all the forecast systems are equally weighted.
- Multi-model-2 was built by pooling all members from all the forecast systems together in a single distribution. This approach was followed for creating both deterministic and probabilistic products from this unified distribution. Therefore, the weight of each forecast system is proportional to the number of members it contributes to the ensemble (see Table 1).

- Multi-model-1-calib was built by using the same approach as multi-model-1 but by previously calibrating the simulations with the method used in Doblas-Reyes et al. (2005) in leave-one-out cross-validation mode independently for each forecast system.
- Multi-model-2-calib was built by using the same approach as multi-model-2 but by previously calibrating the simulations with the method used in Doblas-Reyes et al. (2005) in leave-one-out cross-validation mode independently for each forecast system.

In this study, the anomaly correlation coefficient (ACC; Wilks 2011), the root-mean-square error skill score (RMSSS; Murphy 1988), and the ranked probability skill score (RPSS; Wilks 2011) have been used to evaluate the skill of the forecast systems. The ACC, which measures the linear relationship between two time series, has been used for evaluating the quality of the deterministic products. To do so, the ACC has been computed, at each grid point, between the simulated and observed anomalies' time series. The ACC ranges from −1 to 1. A value of 1 indicates a perfect forecast, while values near zero or negative indicate a forecast with no skill. The RMSSS has also been used to evaluate the quality of the deterministic forecasts. The RMSSS is based on the root-mean-square error (RMSE; Murphy 1988) and estimates the improvement or worsening in the magnitude of the forecast's errors compared to a reference forecast's errors. If the RMSSS is greater than zero, it indicates that the skill of the forecasts is higher than that of the reference forecast. In contrast, negative skill score values mean that the reference forecast is more skillful. The RPSS for three categories (the lower, middle, and upper terciles) has been used to evaluate the probabilistic products. The RPSS is based on the ranked probability score (RPS; Wilks 2011) and provides a measure of the improvement (or lack thereof) of the probabilistic forecast with respect to a reference forecast. Analogous to the RMSSS, if the RPSS is positive (negative), it indicates that the skill of the forecasts is higher (lower) than that of the reference forecast. The ACC and RPSS are insensitive to biases in mean and variance, while the RMSSS is sensitive and accounts for the signal's amplitude.

Four different reference forecasts have been used as benchmarks to compute the RMSSS and RPSS: the climatological forecast (defined as the equiprobable forecast; i.e., probability of 33.33% for each tercile category) is used when assessing the quality of the multi-model approaches; the individual forecast systems are used to assess the benefits and drawbacks of using a multi-model forecast; the historical simulations have been used to estimate the impact of the system initialization; and, finally, a smaller number of simulations has been compared to the complete multi-model ensemble to measure the impact that the ensemble size has on the multi-model skill.

The statistical significance has been tested to check for sampling errors and exclude random effects at the 95% confidence level. For the ACC, a two-sided t test has been used to assess whether the value is significantly different from zero (Wilks 2011). For the ACC differences, the Fisher's z-transformed correlations divided by the standard error of the difference have been used (Wilks 2011). An effective number of degrees of freedom has been used to avoid overestimating the actual significance of the ACC and ACC differences due to the autocorrelation of the time series (von Storch and Zwiers 1999). It should be noted that the traditional significance test used for the ACC differences has been shown to be biased toward indicating no difference in skill (DelSole and Tippett 2014), making the test too conservative and the threshold to reach the significance higher. The significance of the RMSSS and RPSS has been assessed by applying the random walk test (DelSole and Tippett 2016) to the RMSE and RPS time series, respectively, in order to test whether the number of times that the forecast is better or worse than the reference forecast is significant at the 95% confidence level. Therefore, although the skill score is small (or large), the test considers one forecast significantly better (or not) than a reference forecast depending on the number of times that one forecast has provided better predictions than the other forecast, regardless of the actual skill score value.

When measuring the impact of the forecast system initialization, the residual correlation has also been used (in addition to the ACC differences, RMSSS, and RPSS). The residual correlation aims to assess whether the decadal predictions capture any of the observed variability that is not already captured by the historical simulations (Smith et al. 2019; Borchert et al. 2021; Mahmood et al. 2021). The procedure is as follows: the residuals of the decadal predictions' ensemble mean and observations are computed by linearly regressing out the historical simulations' ensemble mean from the decadal predictions' ensemble mean and observations, respectively. Then, the residual correlation is computed as the correlation between both residuals. Positive values of the residual correlation indicate that the decadal predictions capture more observed variability than the historical simulations, while negative values mean that the historical simulations capture more. The significance of the residual correlation is computed as the ACC significance, also taking into account the effective degrees of freedom.

4. Results and discussion

a. *Skill and comparison of multi-model approaches*

In this section, the quality of the DCPP multi-model ensemble predictions is evaluated for near-surface air temperature, precipitation, the AMV index, and the GSAT anomalies. Also, it is assessed whether there are differences in the skill between the different approaches to construct a multi-model.

Maps of ACC obtained for the four multi-model approaches show, in general, similar skill and significance for temperature (first column in Fig. 1; see Fig. S1 in the online supplemental material for ACC differences between the multi-model approaches). For precipitation, lower skill and smaller areas with significant positive skill are found for the calibrated multi-models compared to the noncalibrated (cf. Figs. 1f,h with Figs. 1b,d, respectively). The lower skill of the calibrated multi-models in predicting precipitation can also be

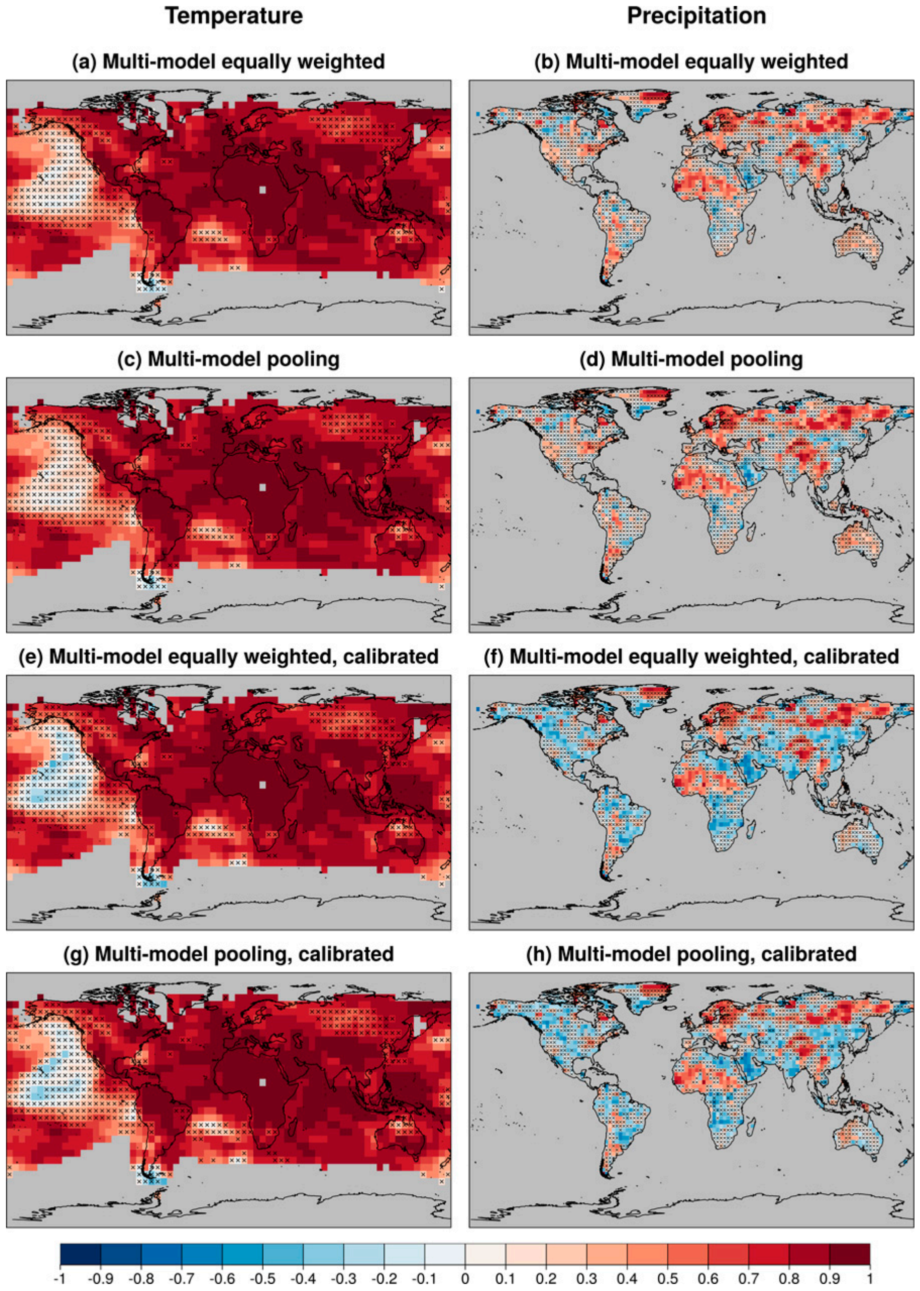

Temperature
Precipitation
(a) Multi-model equally weighted
(b) Multi-model equally weighted
(c) Multi-model pooling
(d) Multi-model pooling
(e) Multi-model equally weighted, calibrated
(f) Multi-model equally weighted, calibrated
(g) Multi-model pooling, calibrated
(h) Multi-model pooling, calibrated
-1 -0.9 -0.8 -0.7 -0.6 -0.5 -0.4 -0.3 -0.2 -0.1 0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

FIG. 1. Maps of ACC obtained with the different multi-model approaches (see section 3) for the forecast years 1–5 for the (left) surface air temperature and (right) precipitation. The ACC has been computed over the 1961–2014 period (start dates: 1960–2009) for each individual grid point. The reference period for the computation of anomalies is 1981–2010. The reference datasets used for the near-surface air temperature and precipitation are the GHCNv4 and the GPCC datasets, respectively. Crosses indicate that the values are not statistically significant at the 95% level using a two-sided *t* test accounting for autocorrelation.

seen in Figs. S1f and S1h, where the red colors indicate a better performance of the noncalibrated multi-model, especially over regions of North America and North Africa, where the differences are significant. It should be noted that the main differences between the multi-model-1 and multi-model-1-calib compared to the multi-model-2 and multi-model-2-calib are due to the different ensemble sizes (the forecast systems with a larger ensemble size contribute more to the latter ones). For temperature, the ACC is generally high and significant. However, there are exceptions like the northeast Pacific Ocean, areas of the South Atlantic Ocean, western Canada, northern Australia, and northern Asia. A similar pattern of the ACC map was found by Smith et al. (2019) using the CMIP5 (Taylor et al. 2012) multi-model ensemble for the forecast years 2–9. For precipitation, the significant positive ACC is limited to some regions of central and northern

Europe, western and central Africa, and northern and central Asia [results in line with Sheen et al. (2017), Smith et al. (2019), and Solaraju-Murali et al. (2019)]. In addition, the multi-model ACC is significantly negative for precipitation in some areas, especially for the calibrated multi-model forecasts, which may be due to cross-validation and to the statistical assumptions of the calibration method (e.g., normal distribution). Also, the calibration method is designed to decrease the forecast error and improve the reliability of the probabilistic forecasts, so the reduction in ACC does not come as a surprise. However, most of the regions with significant skill are common between the four different approaches.

When evaluating the simulations against the second reference dataset (Fig. S2), there is a general agreement with results for temperature in Fig. 1 in the Northern Hemisphere, where there are high-quality observations. However, discrepancies are found over the Arabian Sea, the Philippine Sea, the Indian Peninsula, and the western part of the North Atlantic Ocean, where the correlation obtained against the JRA-55 is not significantly positive (whereas it is significant when using the GHCNv4 as the reference dataset). The results over the Southern Hemisphere are more difficult to compare due to the low coverage of the GHCNv4 dataset. However, for the regions that are covered by both reference datasets, a lower agreement has been found between ACC values in Fig. 1 and Fig. S2 than the one obtained for the Northern Hemisphere, finding a general lower correlation when using the JRA-55 (e.g., over the southern African and Australian coasts). This might be indicative of the quality of reanalyses being affected by fewer observations available for assimilation in the Southern Hemisphere. For precipitation, there are high discrepancies over all the continental regions (the observational uncertainty cannot be addressed over the oceans since the GPCC dataset only provides data over land regions). The only regions where there is an agreement between the ACC values obtained with both reference datasets are areas over Greenland, parts of central and northern Asia, and western Africa.

Maps of RMSSS and RPSS (Figs. 2 and 3, respectively) show similar skill (and significance) for the different multi-model approaches, with spatial patterns similar to those obtained for the ACC (Fig. 1). Also, the patterns of the maps of both skill scores are very similar between them, being the RPSS generally higher than the RMSSS. For temperature, significant positive skill scores (which indicates an added value of the decadal predictions with respect to the climatological forecast) are found, in general, over the same areas as those where ACC was significantly positive. Exceptions for the RPSS are northern Asia (where there is a larger number of significant grid points than in the ACC maps) and the southern part of the Indian Ocean (where the RPSS is not significant). In the Pacific Ocean, there are some grid points where the climatological forecast is significantly better than the multi-model forecasts when measured with both the RMSSS and RPSS. For precipitation, the skill score values are much lower than for temperature, only being significantly positive over limited regions of Africa and Asia, indicating that there is not a significant improvement of the decadal predictions with respect to the climatological forecast over most of the regions. As for temperature, there are also grid points where the climatological forecast is significantly better than the multi-models. Besides, there are no high differences between the RMSSS and RPSS obtained with the first two multi-model approaches (i.e., equally weighted and pooling multi-models; Figs. S3 and S4). Lledó et al. (2020) also found that these two multi-model approaches provide almost identical performance in predicting the Euro-Atlantic teleconnection indices such as the North Atlantic Oscillation (NAO; Wanner et al. 2001; Athanasiadis et al. 2017) in seasonal forecasting. With respect to the calibrated multi-model ensembles, the calibration improves the temperature forecast skill over several areas when measured by the RMSSS (Figs. S3e,g). In contrast, it does not improve the skill when measured by the ACC and RPSS (Figs. S1 and S4).

Comparing the RMSSS and RPSS maps with those obtained using the second reference dataset (Figs. S5 and S6, respectively), there is a general agreement for temperature over the continental areas of the Northern Hemisphere, while large differences are found in regions like South America, southern Africa, the Indian Ocean, the western part of the Pacific Ocean, and Australia. For precipitation, large differences are found between the skill estimates with the different reference datasets (only the continental areas are assessed since the GPCC dataset does not provide data over the oceans), especially over Africa, Europe, and Asia, where the climatological forecast is significantly better than the decadal predictions over most of the regions when using the JRA-55, matching the differences seen with the ACC when using both reference datasets.

As done for temperature and precipitation, the deterministic and probabilistic skills in predicting the AMV and GSAT variability are also assessed (Fig. 4). The results obtained with the DCPP multi-model forecasts show a significant positive ACC, RMSSS, and RPSS for both indices in all the multi-model approaches. The main difference between the approaches is the ensemble spread, which is higher for the multi-model-2 and multi-model-2-calib as they are built with all the members instead of the ensemble means. Besides, the ensemble spread is adjusted in the calibrated multi-models, increasing the reliability of the products and matching the spread with the mean error of the forecasts (Doblas-Reyes et al. 2005), although that metric is not considered in this study. The HIST multi-model ensemble, which also shows significant skill in predicting both indices, is compared to the DCPP multi-model in section 4c to estimate the impact of the system initialization.

The skillful prediction of the AMV index by the decadal predictions is in agreement with previous results in the literature (e.g., García-Serrano et al. 2015; Si et al. 2019; Bilbao et al. 2021). Regarding the GSAT time series, although the positive skill is significant, there is an overestimation of the warming during the most recent period after about year 2000 when observed temperatures show a so-called hiatus (e.g., Fyfe et al. 2016). This overestimation has already been documented for both the CMIP5 and CMIP6 simulations (e.g., Douville et al. 2015; Papalexiou et al. 2020; Tokarska et al. 2020; Wang et al. 2021). This overestimation is partially corrected in the calibrated multi-models, showing a trend more similar to observations (Figs. 4f,h). On the other hand, the cost of applying the

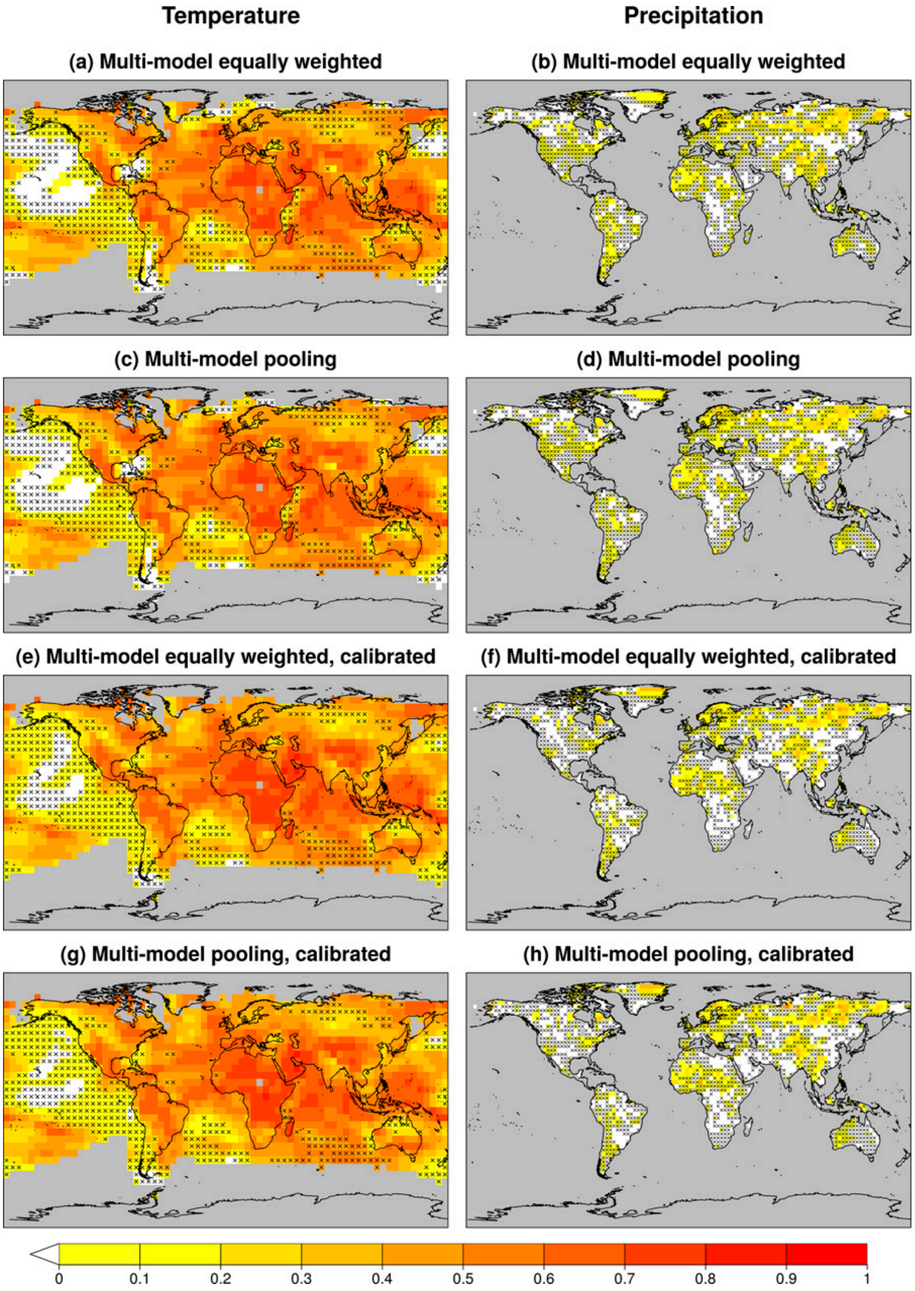

Temperature
Precipitation
(a) Multi-model equally weighted
(b) Multi-model equally weighted
(c) Multi-model pooling
(d) Multi-model pooling
(e) Multi-model equally weighted, calibrated
(f) Multi-model equally weighted, calibrated
(g) Multi-model pooling, calibrated
(h) Multi-model pooling, calibrated
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

FIG. 2. Maps of RMSSS obtained with the different multi-model approaches (see section 3) for the forecast years 1–5 for (left) near-surface air temperature and (right) precipitation using the climatology as the reference forecast. The skill scores have been computed over the 1961–2014 period (start dates: 1960–2009). The reference period for the computation of anomalies is 1981–2010. The reference datasets used for the surface air temperature and precipitation are the GHCNv4 and the GPCC datasets, respectively. Crosses indicate that the decadal predictions do not provide significantly better or worse predictions than the climatological forecast at the 95% confidence level based on a random walk test.

calibration is a worse forecast in the 1960s and 1970s by the calibrated multi-models, when the error is higher.

b. Multi-model ensemble compared to individual forecast systems

For the subsequent analyses, only the multi-model-1 approach is used. This approach has been built by averaging the ensemble means (probabilities) from the individual forecast systems for the deterministic (probabilistic) forecast products. The reason for choosing this approach is that all the forecast systems are equally weighted, while the forecast systems with a larger ensemble size have more weight in the multi-model-2 and multi-model-2-calib approaches. Also, the multi-model-1 is chosen instead of the multi-model-1-calib for

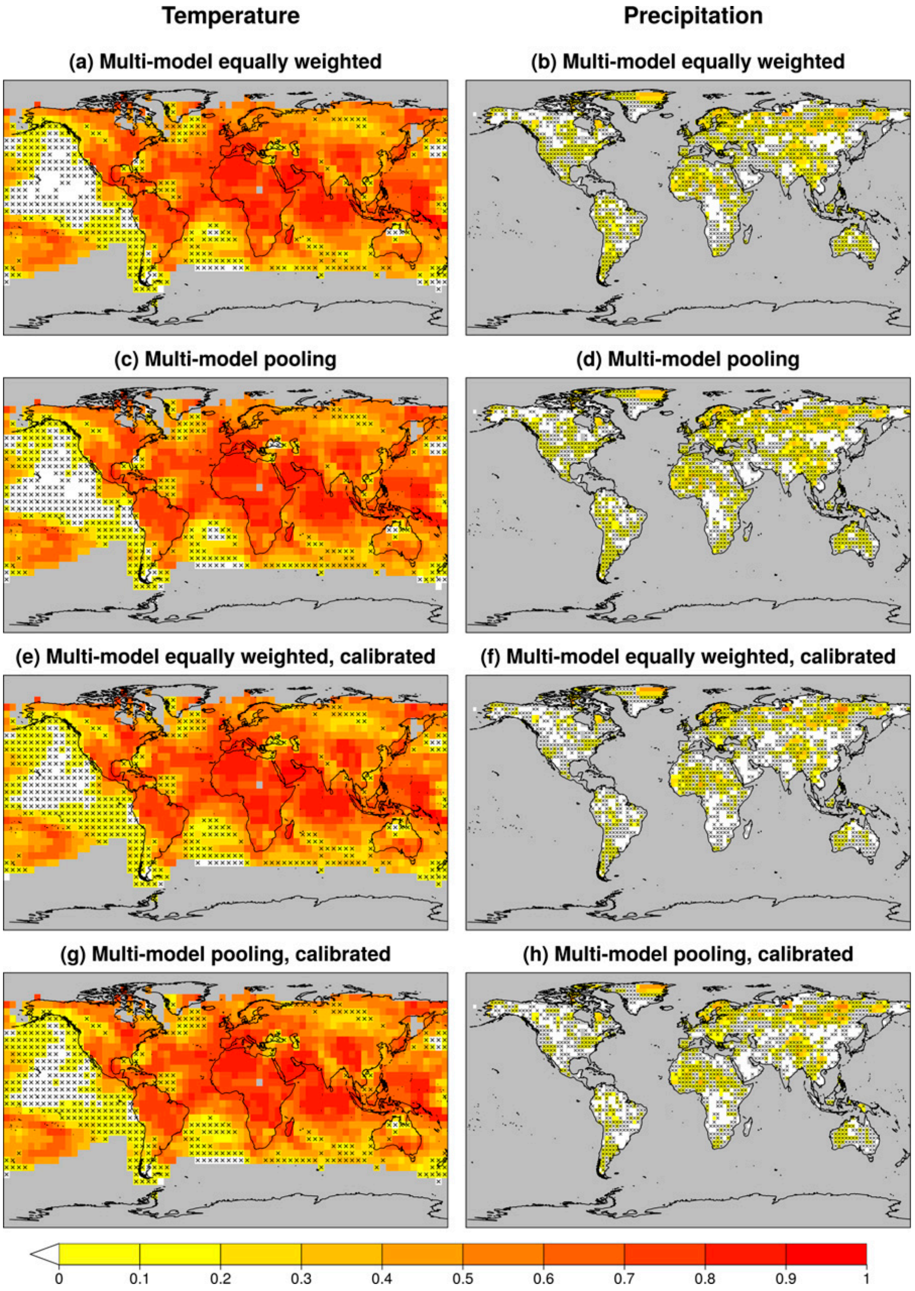

Temperature
Precipitation
(a) Multi-model equally weighted
(b) Multi-model equally weighted
(c) Multi-model pooling
(d) Multi-model pooling
(e) Multi-model equally weighted, calibrated
(f) Multi-model equally weighted, calibrated
(g) Multi-model pooling, calibrated
(h) Multi-model pooling, calibrated
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

FIG. 3. Maps of RPSS for three categories obtained with the different multi-model approaches (see section 3) for the forecast years 1–5 for (left) near-surface air temperature and (right) precipitation using the climatology as the reference forecast. The skill scores have been computed over the 1961–2014 period (start dates: 1960–2009). The reference period for the computation of the thresholds between categories is 1981–2010. The reference datasets used for the surface air temperature and precipitation are the GHCNv4 and the GPCC datasets, respectively. Crosses indicate that the decadal predictions do not provide significantly better or worse predictions than the climatological forecast at the 95% confidence level based on a random walk test.

a fair comparison with the individual forecast systems, which have not been calibrated. Hence, hereafter the multi-model-1 approach is referred to simply as the multi-model.

To assess the advantages and disadvantages of using a multi-model ensemble instead of the simulations from a single forecast system, the skill estimates for the deterministic and probabilistic forecast have been computed for all the systems for temperature and precipitation (Figs. S7–S18). Then, the multi-model is compared to the forecast systems that provide the maximum and median skill values for each grid point (Fig. 5). For the comparison, the ACC differences between the multi-model and the forecast systems that present the maximum and the median skill are directly computed for each grid point. Similarly, the RMSSS and RPSS of the multi-

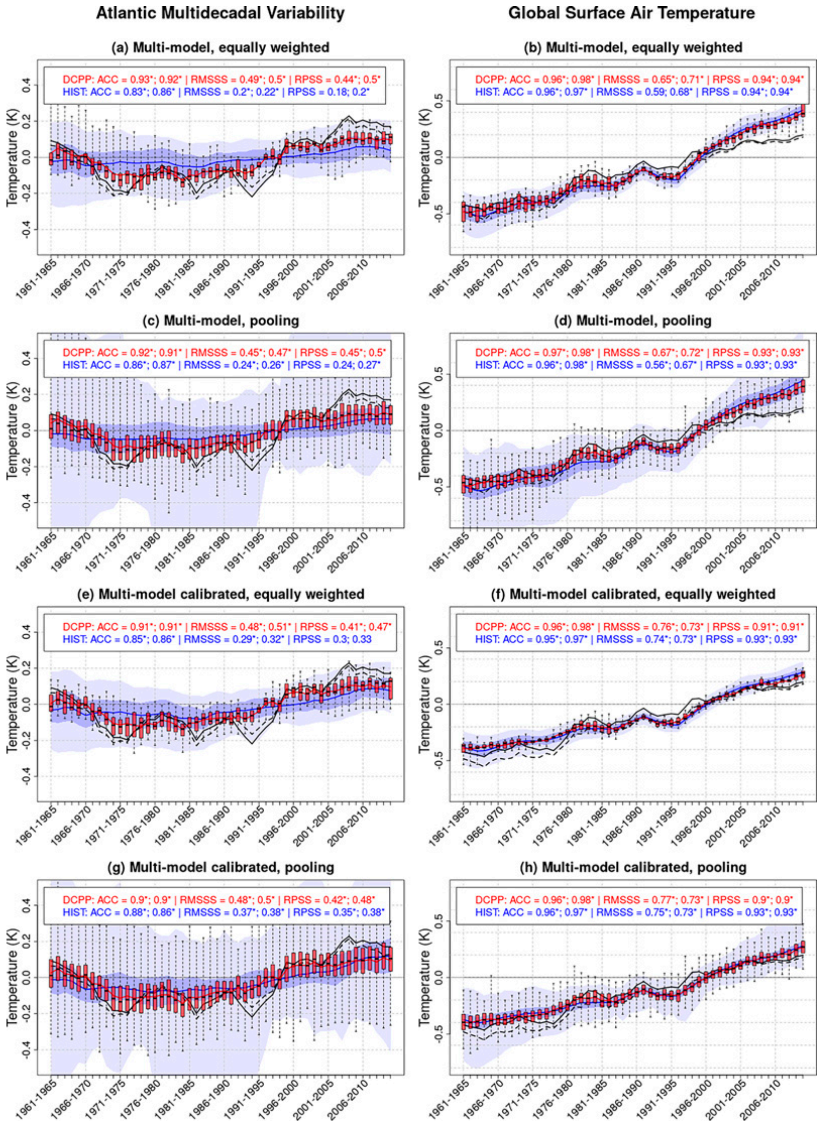


FIG. 4. (left) AMV index and (right) GSAT anomalies obtained with the multi-model approaches for the DCPP (forecast years 1–5) and HIST ensembles. The historical simulations are shown in blue (dark shading contains the values between the 25th and 75th percentiles, while light shading contains the values between those percentiles and the minimum/maximum values) and the decadal predictions in red (boxes contain the values between the 25th and 75th percentiles, while the whiskers contain the values between those percentiles and the minimum/maximum values). In the legend, the ACC, RMSSS, and RPSS are shown for both decadal predictions (in red) and historical simulations (in blue) over 1961–2014 (start dates: 1960–2009). The reference period for the computation of anomalies and thresholds between categories is 1981–2010. The reference datasets used for the AMV index are the GISTEMPv4 (gray solid lines) and the HadCRUT4 (gray dashed lines), while JRA-55 (gray solid lines) and GHCNv4 (gray dashed lines) are used for the GSAT anomalies. The skill measures are shown for both reference datasets in the legend of each panel: the first value corresponds to the GISTEMPv4 (JRA-55) dataset and the second value to the HadCRUT4 (GHCNv4) dataset for the AMV (GSAT). A star next to an ACC estimate indicates that the skill is statistically significant at the 95% confidence level using a two-sided *t* test accounting for autocorrelation, while a star next to an RMSSS or RPSS value indicates that the simulations provide significantly better or worse predictions than the climatological forecast at the 95% confidence level based on a random walk test.

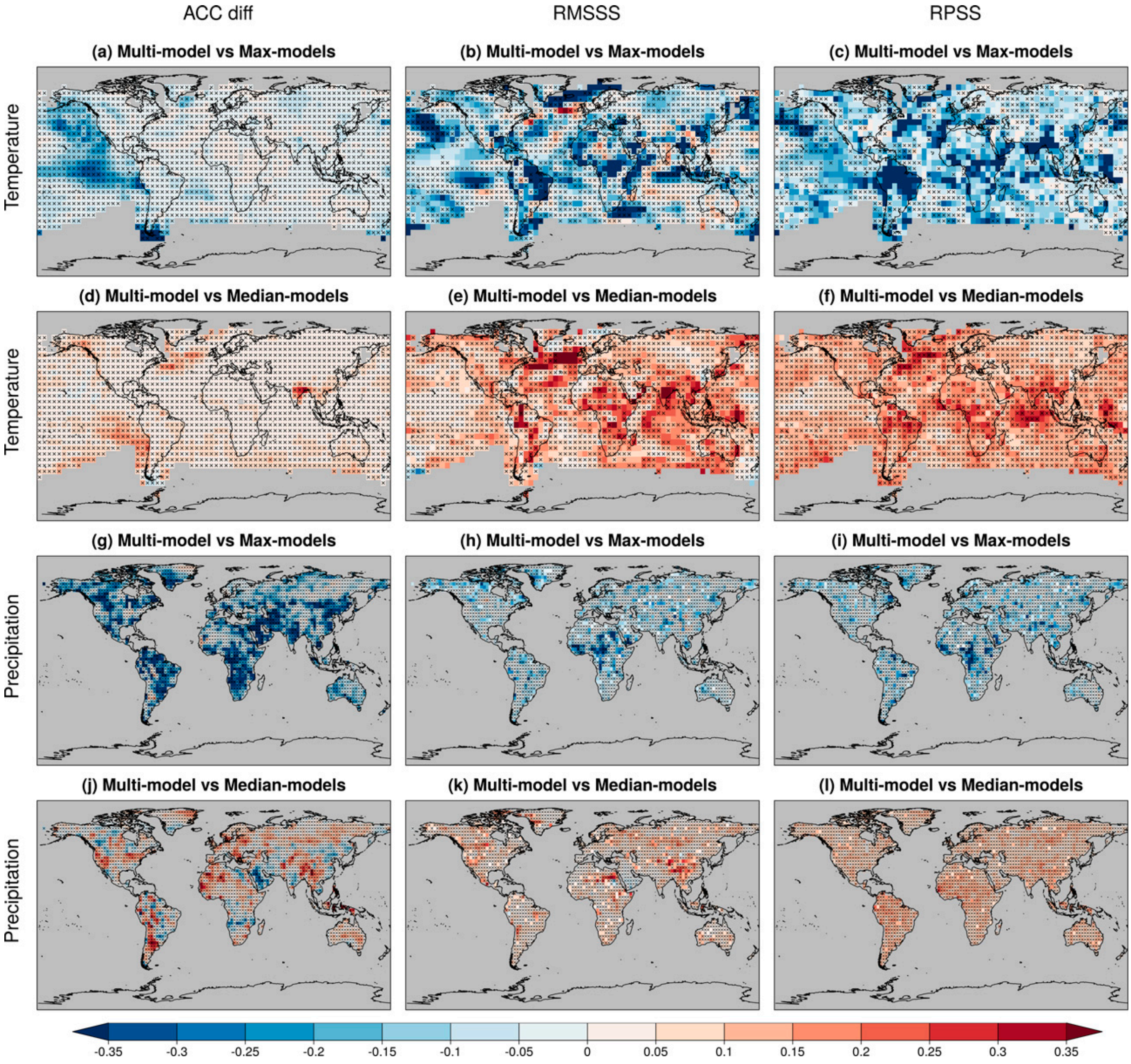

ACC diff
RMSSS
RPSS
(a) Multi-model vs Max-models
(b) Multi-model vs Max-models
(c) Multi-model vs Max-models
Temperature
(d) Multi-model vs Median-models
(e) Multi-model vs Median-models
(f) Multi-model vs Median-models
Temperature
(g) Multi-model vs Max-models
(h) Multi-model vs Max-models
(i) Multi-model vs Max-models
Precipitation
(j) Multi-model vs Median-models
(k) Multi-model vs Median-models
(l) Multi-model vs Median-models
Precipitation
-0.35
-0.3
-0.25
-0.2
-0.15
-0.1
-0.05
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35

FIG. 5. (left) Differences between the ACC obtained with the multi-model and the forecast systems that provide the maximum and median skill, (center) RMSSS of the multi-model using the forecast systems that present the maximum and median skill score as the reference forecast, and (right) RPSS obtained with the multi-model using the forecast systems that provide the maximum and median skill score as the reference forecast. The results are shown for the (a)–(f) near-surface air temperature and (g)–(l) precipitation for the forecast years 1–5. The skill estimates have been computed over the 1961–2014 period (start dates: 1960–2009). The reference period for the computation of the climatology and thresholds between categories is 1981–2010. The reference datasets used for the near-surface air temperature and precipitation are, respectively, the GHCNv4 and the GPCC datasets. For the ACC differences, crosses indicate that the difference is not statistically significant at the 95% confidence level using the two-sided *t* test accounting for autocorrelation. For the RMSSS and RPSS, crosses indicate that the multi-model does not provide significantly better or worse predictions than the best/median forecast system at the 95% confidence level based on a random walk test.

model are calculated using the forecast systems that present the maximum and the median skill as the reference forecasts for each grid point. The aim of this comparison is twofold. On the one hand, comparing the multi-model against the forecast system that provides the highest skill is useful to estimate how much skill is lost when using a multi-model (whose skill may be degraded due to a lower skill of some systems) instead of the best forecast system. On the other hand, the analogous comparison using the forecast system that provides the median skill informs on whether the multi-model provides better predictions than, at least, 50% of the forecast systems. In such a case, using a multi-model might be a worthwhile choice because the forecast products could always be generated using the multi-model ensemble, without having to select the best system for each particular region, forecast period, and variable.

The comparison of the deterministic skill for temperature measured with the ACC (Fig. 5a) shows that the maximum value obtained with the individual forecast systems is similar to that obtained with the multi-model over most of the regions, except for the eastern tropical Pacific Ocean and some regions of the Southern Ocean (only visible with the second reference dataset in Fig. S19a), where the negative differences are the largest (i.e., where the best single system is better than the multi-model). For the case of the eastern tropical Pacific Ocean, the negative differences are due to the higher skill present in the MPI-ESM1.2-LR system than in the multi-model ensemble (Figs. S7l and S7a, respectively) over this region. The skill of this forecast system over the eastern tropical Pacific region indicates that it may capture some sea surface temperature variability at lower frequencies than ENSO. For instance, it might partially capture variations related to the interdecadal Pacific oscillation (IPO; Power et al. 1999), which varies at longer time scales. The map is noisier when using the RMSSS for comparing the multi-model and the best forecast system (Fig. 5b). There are some regions where the multi-model performs better than the best forecast system. However, they are very limited (the highest positive skill scores are found over a few points in the North Atlantic Ocean). In contrast, the best forecast system outperforms the multi-model over most of the globe, finding significant negative skill scores over large areas like regions of central Africa, South America, the Arctic Ocean, and the Pacific Ocean. Regarding the comparison of the probabilistic skill (Fig. 5c), the best system is significantly better than the multi-model over several regions, especially over South America, Africa, and southern Asia. For precipitation, larger negative ACC differences are found over most regions of the globe, indicating a better performance of the best forecast system (Fig. 5d), particularly when using the second reference dataset (Fig. S19d). This better performance of the best forecast system is less evident when estimated with the RMSSS and RPSS, with the skill scores close to zero. Some exceptions are central Africa and a few grid points over North America and Asia, where there are significant negative values of both skill scores.

Although the best single forecast system provides higher skill than the multi-model over most of the regions, it is worth keeping in mind that the best system would have to be determined for each particular region, variable, and forecast period to reach the highest possible skill, complicating the generation of forecast products. Besides, the skill estimates obtained with the multi-model are higher than those obtained with the median skill of the single forecast systems over most of the globe for both variables (although only a few areas show a significant ACC difference or RPSS). For temperature, the areas with the largest benefit of using a multi-model forecast compared to the median of the single systems are the North Atlantic Ocean, central Africa, the Indian Ocean, and the Indian Peninsula for the deterministic products (Figs. 5d,e). The ACC differences were not found to be statistically significant, whereas the positive RMSSS are. Concerning the probabilistic products, the same regions as for RMSSS show an added value of using the multi-model, although with few points that present significance (Fig. 5f). For precipitation, the sign of the ACC differences highly depends on the region, and differences are significant only over a few marginal points (Fig. 5j). The RMSSS and RPSS are positive almost everywhere (although with the significance limited to a few points), indicating a higher skill of the multi-model ensemble (Figs. 5k,l). Then, although the best forecast system provides higher skill than the multi-model over most of the regions, the multi-model presents higher skill than the median of the single systems, meaning that there is more than 50% of probability of choosing worse predictions if a random individual system is used, and without the need of having to select the best system for each particular region, forecast period, and variable (Mishra et al. 2018; Hemri et al. 2020). On the other hand, in a climate services context, the best forecast system or multi-model approach could be selected to issue the best possible predictions over a particular region, variable and forecast period (Fig. S20). For instance, the multi-model precipitation forecast for the Arabian Peninsula is worse than the median value obtained with the individual systems (as shown for the ACC differences in Fig. 5j). Thus, when creating a climate service for this region, it would be worth selecting the forecast system that presents the highest skill or creating a multi-model forecast from a subsample of systems if one could be shown to have higher skill than this single system.

The comparison between the multi-model and the individual forecast systems has also been performed for the AMV and GSAT time series. For the AMV, most of the forecast systems present significantly positive ACC, RMSSS, and RPSS values (Fig. S21) with the exceptions of the BCC-CSM2-MR, CanESM5, and MPI-ESM1.2-LR systems. For the GSAT anomalies, all the forecast systems show significantly positive ACC and RPSS (Fig. S22). Still, similar to the multi-model approaches (Fig. 4), all the forecast systems overestimate the warming trend during the most recent part of the evaluation period.

To compare the skill of the multi-model and the individual forecast systems, the ACC differences between the multi-model and the single systems and the RMSSS and RPSS of the multi-model using the forecast systems as the reference forecast have been computed and displayed in Fig. 6. For the AMV, the ACC differences, RMSSS and RPSS are positive, indicating a better performance of the multi-model (except the RMSSS obtained using the MRI-ESM2-0 system as reference, which is negative). However, the comparison is only significant with respect to the BCC-CSM2-MR, CanESM5, and MPI-ESM1.2-LR systems for the ACC difference; with respect to BCC-CSM2-MR, CanESM5, EC-Earth3-i1, IPSL-CM6A-LR, MPI-ESM1.2-LR, and NorCPM1 for the RMSSS; and with respect to the MRI-ESM2-0 system for the RPSS. The same results are found when using the second reference dataset (Fig. S23), except for the RMSSS, which is only significant when using the BCC-CSM2-MR, CanESM5, and MPI-ESM1.2-LR systems as reference forecasts. For the GSAT, low (and not significant) ACC differences are found between the multi-model and the forecast systems. The

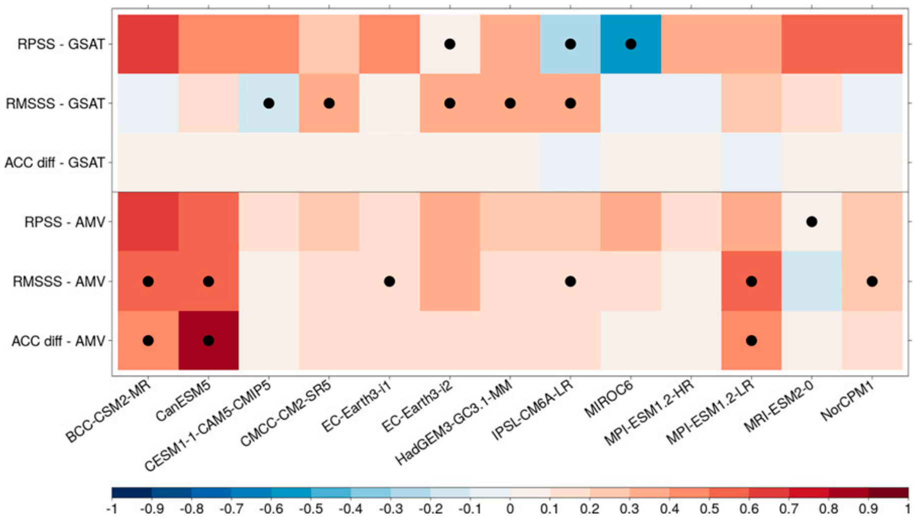


FIG. 6. Differences between the ACC obtained with the multi-model and the individual forecast systems, and RMSSS and RPSS obtained with the multi-model using the forecast systems as the reference forecasts. The differences and skill scores are shown for the AMV and GSAT indices for the forecast years 1–5. The skill estimates have been computed over the 1961–2014 period (start dates: 1960–2009). The reference period for the computation of the climatology and thresholds between categories is 1981–2010. The reference datasets used for the AMV and GSAT indices are, respectively, the GISTEMPv4 and JRA-55 datasets. For the ACC, dots indicate that the differences are statistically significant at the 95% level using a two-sided t test accounting for autocorrelation. For the RMSSS and RPSS, dots indicate that the multi-model provides significantly better or worse predictions than the forecast systems at the 95% confidence level based on a random walk test.

RMSSS shows an added value of using a multi-model instead of any individual systems, which is significant when using the EC-Earth3-i1, EC-Earth3-i2, HadGEM3-GC3.1-MM, IPSL-CM6A-LR, MIROC6, and MPI-ESM1.2-HR systems as reference. The comparison of the probabilistic skill shows that the multi-model is significantly better than the EC-Earth3-i2 system, while it is significantly worse than the IPSL-CM6A-LR and MIROC6 systems. The same results are found for the ACC differences and RPSS when using the second reference dataset (Fig. S23). For the RMSSS, the skill scores are significant for the CESM1-1-CAM5-CMIP5 (negative in this case; i.e., outperforming the multi-model), CMCC-CM2-SR5, EC-Earth3-i2, HadGEM3-GC3.1-MM, and IPSL-CM6A-LR systems. The lack of significance for the ACC differences may be due to the low effective number of degrees of freedom that the time series have due to their high autocorrelation (10.7 and 5.1 for the observed AMV and GSAT, respectively). It should be noted that the random walk test (used to test the RMSSS and RPSS significance) is based on the significance of the number of years that one forecast overcomes the reference forecast. Thus, although the RMSSS and RPSS are small (or large), the test considers them as significant (or not) depending on the number of years that one forecast provides better predictions than the other one, regardless of the actual value of the skill score.

c. Impact of forecast system initialization

The multi-model ensemble built with decadal predictions is compared to the one built with historical simulations. This comparison is made in order to assess the impact of the forecast system initialization toward the observed climate state, while the forcings are the same between the DCPP and HIST multi-models. Figure 7 shows the ACC differences between the decadal predictions and the historical simulations, the residual ACC (i.e., the correlation between the residuals of the DCPP ensemble mean and observations once the HIST ensemble mean has been linearly regressed out from each; Smith et al. 2019), and the RMSSS and RPSS of the decadal predictions using the historical simulations as the reference forecast for the near-surface air temperature and precipitation.

For the near-surface air temperature, the differences between the ACC obtained with DCPP and HIST do not show a significant impact due to initialization (Fig. 7a), except for some areas over the Southern Ocean (only visible when using the JRA-55 as the reference dataset; Fig. S24a). Although without significance, the North Atlantic Subpolar Gyre region also shows an improvement after initialization. The initialization in this region has been shown to be of high importance for maximizing skill in the area (e.g., Yeager et al. 2018). The low ACC differences may be explained by the high ACC that the historical simulations already show in predicting the near-surface temperature. Note that the significance of the ACC differences may be underestimated due to the bias of the traditional significance test toward showing no significant differences in skill when evaluated against the same reference dataset (DelSole and Tippett 2014).

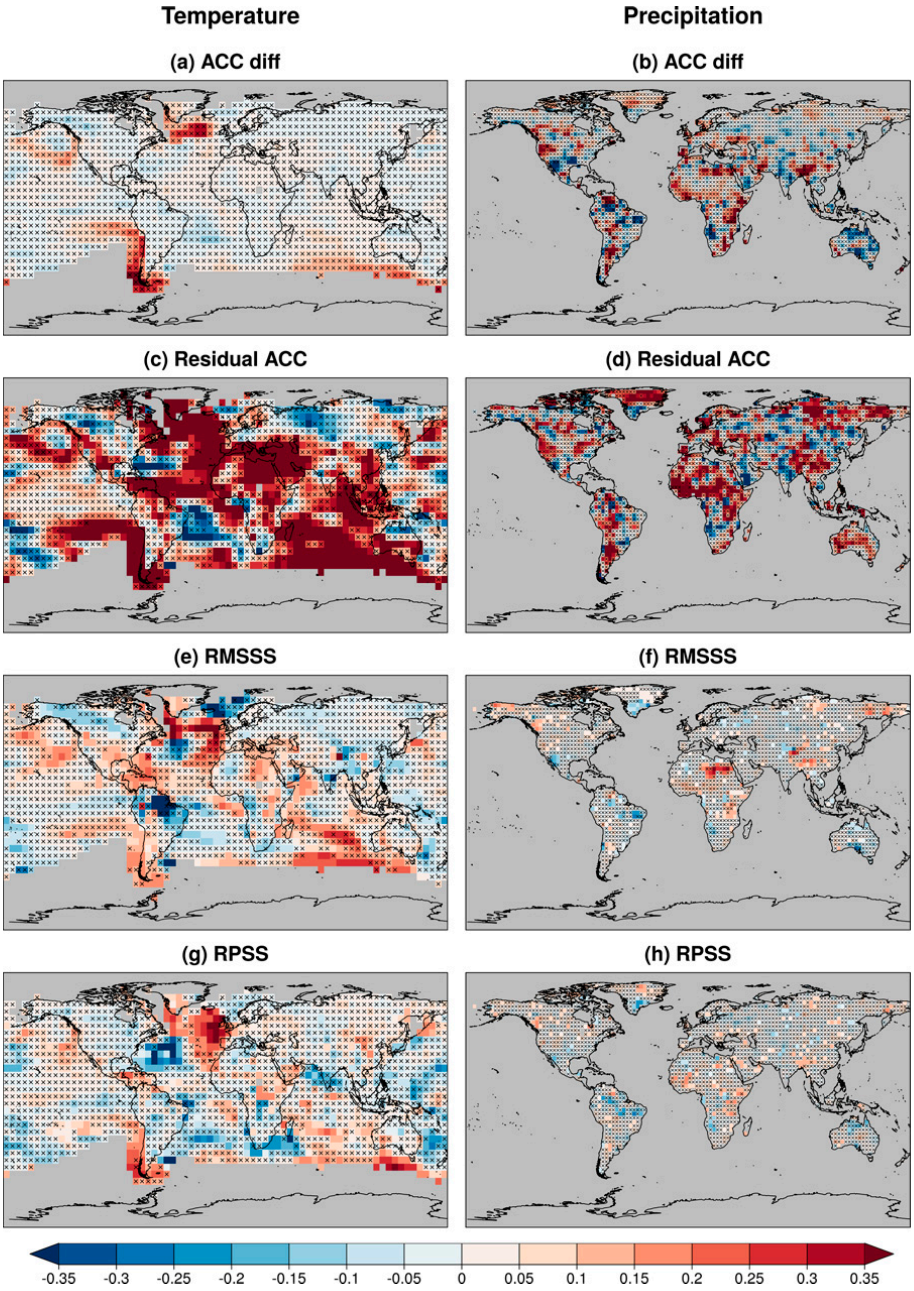


FIG. 7. Maps of (a),(b) ACC differences between the decadal predictions and historical simulations, (c),(d) residual ACC, (e),(f) RMSSS of the decadal predictions using the historical simulations as the reference forecast, and (g),(h) RPSS of the decadal predictions using the historical simulations as the reference forecast. The values are shown for the (left) near-surface air temperature and (right) precipitation for the forecast years 1–5 with the multi-model. The skill estimates have been computed over the 1961–2014 period (start dates: 1960–2009). The reference period for the computation of the climatology and thresholds between categories is 1981–2010. The reference datasets used for the near-surface air temperature and precipitation are, respectively, the GHCNv4 and the GPCC datasets. For the ACC difference and residual ACC maps, crosses indicate that the values are not statistically significant at the 95% level using a two-sided *t* test accounting for autocorrelation. For the RMSSS and RPSS maps, crosses indicate that the decadal predictions do not provide significantly better or worse predictions than the historical simulations at the 95% confidence level based on a random walk test.

Therefore, the residual correlation has also been used for assessing the added value of the forecast system initialization, as proposed by Smith et al. (2019). The residual correlation shows a significant improvement due to initialization over some ocean and land regions. Over the ocean, the highest added value is found over the northeastern and central regions of the Atlantic Ocean, the southwestern corner of the Pacific Ocean, the Southern Ocean, and the Indian Ocean. Over land, significant improvement is found over regions such as North Africa and the Arabian Peninsula (Fig. 7c). By contrast, negative residual correlation is found over other regions like the western part of the North Atlantic Ocean and parts of the South Atlantic Ocean. The skill scores computed using the historical simulations as the reference forecast also show a benefit of initialization over some regions. The RMSSS (Fig. 7b) is significantly positive over regions like the eastern part of the North Atlantic Ocean, Central America, and areas of North America and northern Africa. On the other hand, it shows significantly negative values over regions of South America and the Arctic Ocean. For the probabilistic products, the RPSS (Fig. 7d) shows an improvement of the decadal predictions with respect to the historical simulations over the eastern North Atlantic, the western coast of South America, and the Southern Ocean (see Fig. S24e for the latter). On the contrary, some regions present a worsening of RPSS due to initialization, like the western part of the North Atlantic Ocean.

The maps of ACC differences and residual ACC for precipitation are noisier (Figs. 7b and 7d, respectively), and a positive or negative impact is found depending on the region. For instance, the residual ACC shows a positive impact of initialization over parts of central Africa, South America, and Asia. In contrast, significantly negative residual ACC is found, for example, over the Arabian Peninsula. The RMSSS and RPSS maps for precipitation also shows an improvement or worsening depending on the region, with skill scores close to zero (Figs. 7f,h). The only remarkable region is the northeastern corner of Africa, where a significantly positive RMSSS is found.

The impact that the initialization has on the forecast quality has been addressed in several studies with CMIP5 and CMIP6 simulations (e.g., Doblas-Reyes et al. 2013; Meehl et al. 2014; Caron et al. 2015; Smith et al. 2019; Borchert et al. 2021; Bilbao et al. 2021). The benefit of initialization has been reported mainly for regions over the North Atlantic, Pacific, and Indian Oceans. In particular, Borchert et al. (2021) documented that the skill had improved from CMIP5 to CMIP6 for the subpolar North Atlantic, primarily related to improved forcing, resulting in a smaller added value from initialization in CMIP6. Besides, Smith et al. (2019) also found a significant positive impact of initialization over land regions such as southern Europe and central Africa using the residual correlation methodology, showing a very similar pattern to the one in Fig. 7c.

The comparison of the DCPP and HIST multi-models' skill in predicting the AMV and GSAT time series is shown in Table 3 to assess the impact of initialization. For the AMV index, a low correlation difference is found. However, the residual ACC, RMSSS, and RPSS are significantly positive, indicating an added value due to initialization. In addition, the HIST ensemble mean shows a lower AMV variance, indicating that even if external forcings have an impact on the AMV, the full signal cannot be reproduced without realistic initialization, pointing to the importance of internal variability for its prediction. Besides, the system initialization may correct the model response to external forcings, improving the forecast quality. The better AMV predictions provided by the decadal predictions with respect to the historical simulations have also been shown by García-Serrano et al. (2015), Si et al. (2019), and Bilbao et al. (2021). For the GSAT anomalies, the added value is seen when using the residual ACC and RMSSS, both being significantly positive (except the RMSSS when using GHCNv4 as the reference, which is not significant). The values of the ACC difference and RPSS are very close to zero and not significant. In addition, it can be seen that the overestimation of recent global warming is lower in the DCPP ensemble, although it is still present. This partially corrected overestimation was also found for the CMIP5 multi-model ensemble (Meehl et al. 2014).

d. *Impact of the multi-model ensemble size*

The quality of the ensemble predictions is expected to improve as more members are used (Smith et al. 2019; Athanasiadis et al. 2020). However, in an operational context, not all prediction centers provide their simulations in near–real time for the forecast product generation, thus limiting the ensemble size for actual forecast applications. With the aim of estimating how the skill is impacted by using a smaller ensemble, the skills of the full DCPP ensemble (169 decadal prediction members from 13 forecast systems) and a smaller DCPP subensemble (40 decadal prediction members from 4 forecast systems: the CMCC-CM2-SR5, EC-Earth3-i1, HadGEM3-GC3.1-MM, and MPI-ESM1.2-HR systems). These four forecast systems have been selected to construct the smaller multi-model ensemble because the centers that run their simulations have the capacity to provide timely simulations for an operational multi-model product generation. Besides, the Copernicus Climate Change Service (C3S) operated by the European Centre for Medium-Range Weather Forecasts (ECMWF) has selected these forecast systems for a prototype of climate services for decadal predictions (C3S_34c contract; https://climate.copernicus.eu/c3s34c-prototype-service-decadal-climate-predictions). Therefore, this smaller multi-model is referred to as the C3S_34c multi-model from here.

Figure 8 shows the differences in the ACC between the full and C3S_34c multi-model ensembles, as well as the RMSSS and RPSS of the full ensemble using the C3S_34c ensemble as the reference forecast (i.e., the comparison of the DCPP multi-models with different ensemble sizes: 169 vs 40 members). For temperature, the ACC differences are positive (which indicates a higher quality of the full ensemble; Fig. 8a) over parts of the eastern Pacific Ocean, the North Atlantic Subpolar Gyre, the Indian Peninsula, and the Southern Ocean (the latter only visible when using the second reference dataset; Fig. S25a). However, these differences are not statistically significant, except for a few points over the Southern Ocean.

TABLE 3. ACC difference between the decadal predictions and historical simulations, residual ACC, RMSSS, and RPSS of the DCPP multi-model ensemble using the HIST multi-model ensemble as the reference forecast for the AMV index and GSAT anomalies for the forecast years 1–5. The skill estimates have been computed over the 1961–2014 period (start dates: 1960–2009). The reference period for the computation of the climatology and thresholds between categories is 1981–2010. For the ACC difference and residual ACC, an asterisk indicates it is statistically significant at the 95% level using a two-sided *t* test accounting for autocorrelation. For the RMSSS and RPSS, an asterisk indicates that the decadal predictions provide significantly better or worse predictions than the historical simulations at the 95% confidence level based on a random walk test.

Index (reference dataset)	ACC difference	Residual ACC	RMSSS	RPSS
AMV (GISTEMPv4)	0.1	0.75*	0.37*	0.32*
AMV (HadCRUT4)	0.06	0.69*	0.36*	0.38*
GSAT (JRA-55)	0.01	0.46*	0.16*	−0.06
GSAT (GHCNv4)	0	0.45*	0.08	−0.06

The RMSSS map (Fig. 8b) shows a significantly better quality in terms of errors of the full ensemble over large regions of the North Atlantic Ocean, southern Africa, South America, the northern and central Pacific Ocean, and the Southern Ocean. In contrast, the RMSSS is significantly negative over other regions like the Arabian Peninsula. The results of the comparison of the probabilistic skill (Fig. 8c) reveal a significant benefit of using the full multi-model ensemble over areas of the Pacific Ocean, Africa, South America, and the Southern Ocean. On the other hand, there are also significant RPSS values (i.e., the smaller ensemble provides higher skill) over other regions, particularly over the Arabian Peninsula and parts of North America, Asia, and Australia.

For precipitation, there are some areas where there is a benefit in skill due to using the larger ensemble, for example, over parts of western Australia, North America, and South America (Fig. 8d). In contrast, there are also a few grid points where the skill of the C3S_34c ensemble is higher (e.g., over the western part of South America). However, both positive and negative ACC differences are not significant. With respect to the RMSSS and RPSS (Figs. 8e,f), the benefit of using the full ensemble is generally positive, being statistically significant over some areas of all the continents. In the case of the RMSSS, significantly negative values are also found over limited regions of Africa. Still, it should be noted that most of the areas with a significant benefit of using the full ensemble instead of the C3S_34c ensemble (e.g., the northern part of South America and central Africa) present negative or non-significantly positive RMSSS and RPSS with respect to the climatological forecast (see Figs. 2b and 3b).

5. Summary and conclusions

In this work, a deterministic and probabilistic forecast quality assessment has been performed using all the available decadal predictions from the forecast systems contributing to the DCPP-A component of the CMIP6. The evaluation has been applied over two essential climate variables (near-surface air temperature and precipitation) and two indices (global surface air temperature anomalies and Atlantic multidecadal variability) for the average of the forecast years 1–5. It should be noted that the quality of the products can vary if other variables and forecast periods are considered. The ACC and the RMSSS have been used to evaluate the quality of the deterministic products, while the RPSS has been used to evaluate the probabilistic products. The choice of these metrics is motivated by the fact that they assess different aspects of the forecast quality.

The quality of the forecast products from four different multi-model approaches has been compared. The benefits and drawbacks of using a multi-model ensemble instead of the simulations from a single forecast system have also been documented. In addition, the impact that the system initialization has on the forecast quality has been assessed by comparing the skill of the DCPP and HIST multi-model ensembles. Finally, two multi-model ensembles built with different ensemble sizes have been compared to estimate how the skill is affected due to the limited number of forecast producing centers that can provide timely decadal predictions for the multi-model products generation in an operational context to underpin climate services. All the quality estimates have been computed using two different reference datasets to account for the observational uncertainty, which is particularly high for precipitation and, to a lower extent, for temperature over oceanic regions. Future improvements in observation-based datasets are expected to increase the robustness of the forecast quality assessment as well as to improve the realistic initialization of the predictions.

The skill of the DCPP multi-model ensemble is generally high for near-surface air temperature, particularly over land regions. Compared to temperature, the skill is lower for precipitation and is limited to regions over central Africa, Europe, and Asia. The four multi-model approaches provide, in general, similar skill and significance for temperature. For temperature, some regions like central Africa and the Arabian Peninsula benefit from calibration when measured by the RMSSS (the ACC and RPSS show no or little benefit). For precipitation, the skill of the calibrated multi-models is lower, which might be linked to the assumption of a normal distribution. Also, the multi-model ensembles tend to get less benefit from calibration than the individual forecast systems (Doblas-Reyes et al. 2005; Hemri et al. 2020). With respect to the indices, the main difference between the multi-model approaches is the lower overestimation of global warming by the calibrated multi-models during the last part of the evaluation period. However, the calibrated multi-models show a worse forecast in the 1960s and 1970s for the GSAT anomalies.

The comparison between the quality of the multi-model products and those created with individual forecast systems shows benefits but also drawbacks. On the one hand, using the best forecast system (or multi-model ensemble) for each particular location, variable, and forecast period provides the highest possible quality of the forecast product, which is what Mishra et al. (2018) recommend for seasonal forecasting. On the other hand, using a multi-model provides a higher skill than, at least, the 50% of the forecast systems, without the

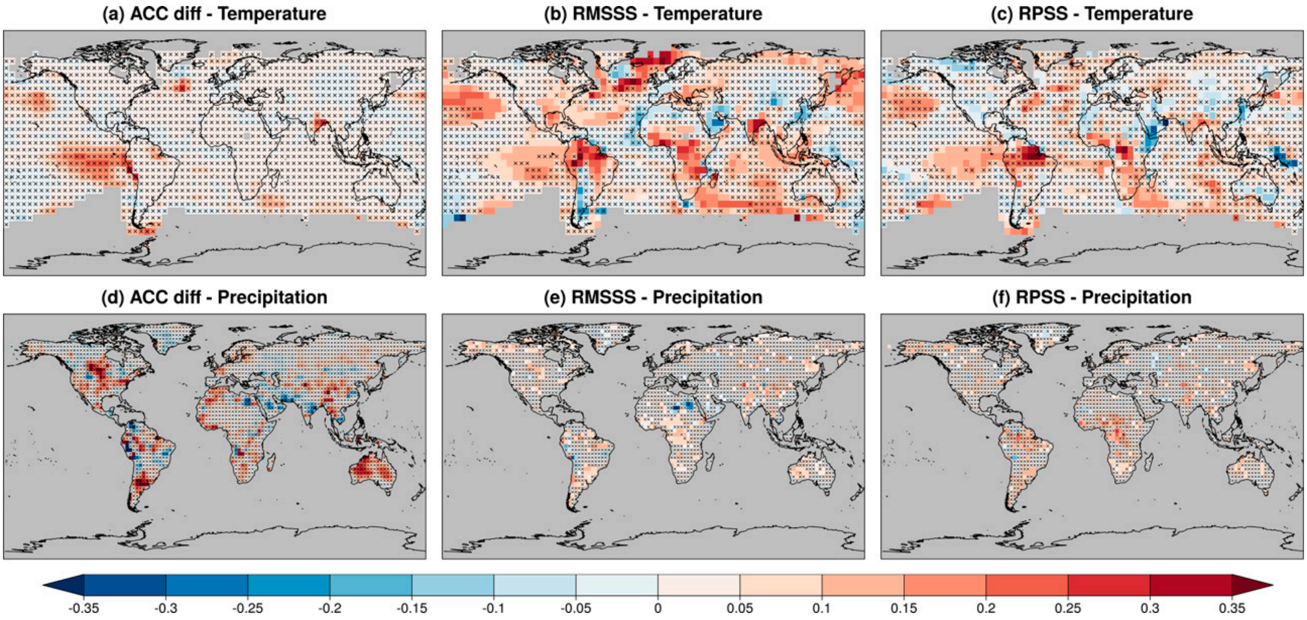


FIG. 8. Maps of (left) ACC differences between the DCPP multi-model and the C3S_34c multi-model, (center) RMSSS of the DCPP multi-model using the C3S_34c multi-model as the reference forecast, and (right) RPSS obtained with the DCPP multi-model using the C3S_34c multi-model as the reference forecast. The DCPP multi-model is built with 169 members from 13 forecast systems, while the C3S_34c multi-model is built with 40 members from four forecast systems (the CMCC-CM2-SR5, EC-Earth3-i1, HadGEM3-GC3.1-MM, and MPI-ESM1.2-HR systems). The differences are shown for the (top) near-surface air temperature and (bottom) precipitation for the forecast years 1–5. The skill estimates have been computed over the 1961–2014 period (start dates: 1960–2009). The reference period for the computation of the climatology and thresholds between categories is 1981–2010. The reference datasets used for the near-surface air temperature and precipitation are, respectively, the GHCNv4 and the GPCC datasets. For the ACC difference maps, crosses indicate that the difference is not statistically significant at the 95% confidence level using a two-sided *t* test accounting for autocorrelation. For the RMSSS and RPSS maps, crosses indicate that the DCPP multi-model does not provide significantly better or worse predictions than the C3S_34c multi-model ensemble at the 95% confidence level based on a random walk test.

need to select the best system for each particular case, which makes an operational forecast product generation more straightforward (Lledó et al. 2020).

The forecast system initialization provides positive or negative impacts depending on the region and variable considered. This dependency might be due to the different quality of assimilated observational data in different regions and variables, and different regions being differently affected by deteriorating effects such as initialization shocks. For temperature, there is a significant added value of the initialization over land (e.g., northern Africa and the Arabian Peninsula) and ocean (e.g., the eastern part of the North Atlantic Ocean and the Indian Ocean) regions. Also, the predictions over the North Atlantic Subpolar Gyre region are improved due to the initialization. The skill in this region is of high importance since low-frequency variability here is suggested to be linked with the skill found for high-latitude blocking and the NAO (Athanasiadis et al. 2020). By contrast, there is a worsening over other areas like the western part of the North Atlantic Ocean. Both the improvement and worsening are more evident for the probabilistic products. For precipitation, the impact is noisier, while the statistical significance is lower than for temperature. The AMV and GSAT predictions are improved due to initialization. In the GSAT time series, the overestimation of recent global warming is higher with the HIST than with the DCPP ensemble, meaning that the initialization partially corrects this issue (also seen in the CMIP5 multi-model ensemble; Meehl et al. 2014).

The differences in the skill of the multi-models built using a different number of forecast systems show, in general, a benefit of using a large ensemble. This result was also found by Smith et al. (2019) for the deterministic forecast of the NAO, which was attributed to the signal-to-noise paradox (Eade et al. 2014; Dunstone et al. 2016; Scaife and Smith 2018). Furthermore, Athanasiadis et al. (2020) also showed the benefit of using a larger ensemble for the high-latitude blocking and NAO predictions. The skill of the precipitation forecasts also shows the benefit of using a large number of simulations over some areas, although it is lower than for temperature. However, the regions with significantly higher skill in the full DCPP ensemble as compared to the subset providing quasi-operational forecasts within C3S_34c are quite limited, and most of them coincide with regions that do not present a significant improvement with respect to the climatological forecast.

In a climate services context, the forecast quality assessment is essential for providing high-quality and reliable forecast products that can be used for decision-making in several sectors. The quality estimates are specific and should be provided with the particular forecast product issued, which varies according to their intended use (Goddard et al. 2013). This study has focused on the prediction of temperature and precipitation for the average of forecast years 1 to 5. However, the transferability of the results is limited and depends on the specific variable, forecast period (one or more years, seasons, or months), and region considered (Sgubin et al. 2021). Thus, the same exercise must be carried out for the specific climate service that aims to be provided.

Acknowledgments. This study has been performed in the framework of the C3S_34c contract (ECMWF/COPERNICUS/2019/C3S_34c_DWD) of the Copernicus Climate Change Service (C3S) operated by the European Centre for Medium-Range Weather Forecasts (ECMWF) and the European Commission H2020 EUCP project (Grant 776613). CDT thanks the funding by the Spanish Ministry for Science and Innovation (FPI PRE2019-088646). MGD is grateful for funding by the Spanish Ministry for the Economy, Industry and Competitiveness grant reference RYC-2017-22964. BSM acknowledges financial support from the Marie Sklodowska-Curie fellowship (Grant 713673) and from a fellowship of La Caixa Foundation (ID 100010434). The authors thank all the modeling groups (see Table 1) for making the simulations available through the Earth System Grid Federation system (ESGF). The authors thank three anonymous reviewers for their valuable comments and suggestions, which improved the manuscript.

Data availability statement. The decadal predictions and historical simulations that support the results of this study are openly available at the Earth System Grid Federation system (ESGF; https://esgf-node.llnl.gov/projects/cmip6/). The data for the GHCNv4 dataset are available at ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/v4/. The data for the JRA-55 are available at https://rda.ucar.edu/datasets/ds628.1/. The data for the GPCC dataset are available at https://opendata.dwd.de/climate_environment/GPCC/html/fulldata-monthly_v2018_doi_download.html. The data for the GISTEMPv4 dataset are available at https://data.giss.nasa.gov/gistemp/. The data for the HadCRUT4 dataset are available at https://crudata.uea.ac.uk/cru/data/temperature/.

REFERENCES

Athanasiadis, P. J., and Coauthors, 2017: A multisystem view of wintertime NAO seasonal predictions. *J. Climate*, **30**, 1461–1475, https://doi.org/10.1175/JCLI-D-16-0153.1.

——, S. Yeager, Y.-O. Kwon, A. Bellucci, D. W. Smith, and S. Tibaldi, 2020: Decadal predictability of North Atlantic blocking and the NAO. *npj Climate Atmos. Sci.*, **3**, 20, https://doi.org/10.1038/s41612-020-0120-6.

Bellucci, A., and Coauthors, 2014: An assessment of a multi-model ensemble of decadal climate predictions. *Climate Dyn.*, **44**, 2787–2806, https://doi.org/10.1007/s00382-014-2164-y.

Bethke, I., and Coauthors, 2021: NorCPM1 and its contribution to CMIP6 DCPP. *Geosci. Model Dev.*, **14**, 7073–7116, https://doi.org/10.5194/gmd-14-7073-2021.

Bilbao, R., and Coauthors, 2021: Assessment of a full-field initialized decadal climate prediction system with the CMIP6 version of EC-Earth. *Earth Syst. Dyn.*, **12**, 173–196, https://doi.org/10.5194/esd-12-173-2021.

Boer, G. J., and Coauthors, 2016: The Decadal Climate Prediction Project (DCPP) contribution to CMIP6. *Geosci. Model Dev.*, **9**, 3751–3777, https://doi.org/10.5194/gmd-9-3751-2016.

Borchert, L. F., M. B. Menary, D. Swingedouw, G. Sgubin, L. Hermanson, and J. Mignot, 2021: Improved decadal predictions of North Atlantic subpolar gyre SST in CMIP6. *Geophys. Res. Lett.*, **48**, e2020GL091307, https://doi.org/10.1029/2020GL091307.

Boucher, O., and Coauthors, 2020: Presentation and evaluation of the IPSL-CM6A-LR climate model. *J. Adv. Model. Earth Syst.*, **12**, e2019MS002010, https://doi.org/10.1029/2019MS002010.

Bruno Soares, M., M. Alexander, and S. Dessai, 2018: Sectoral use of climate information in Europe: A synoptic overview. *Climate Serv.*, **9**, 5–20, https://doi.org/10.1016/j.cliser.2017.06.001.

Caron, L.-P., L. Hermanson, and F. J. Doblas-Reyes, 2015: Multiannual forecasts of Atlantic U.S. tropical cyclone wind damage potential. *Geophys. Res. Lett.*, **42**, 2417–2425, https://doi.org/10.1002/2015GL063303.

——, ——, A. Dobbin, J. Imbers, L. Lledó, and G. A. Vecchi, 2018: How skillful are the multiannual forecasts of Atlantic hurricane activity? *Bull. Amer. Meteor. Soc.*, **99**, 403–413, https://doi.org/10.1175/BAMS-D-17-0025.1.

Cherchi, A., and Coauthors, 2019: Global mean climate and main patterns of variability in the CMCC-CM2 coupled model. *J. Adv. Model. Earth Syst.*, **11**, 185–209, https://doi.org/10.1029/2018MS001369.

Cowtan, K., and Coauthors, 2015: Robust comparison of climate models with observations using blended land air and ocean sea surface temperatures. *Geophys. Res. Lett.*, **42**, 6526–6534, https://doi.org/10.1002/2015GL064888.

DelSole, T., and M. K. Tippett, 2014: Comparing forecast skill. *Mon. Wea. Rev.*, **142**, 4658–4678, https://doi.org/10.1175/MWR-D-14-00045.1.

——, and ——, 2016: Forecast comparison based on random walks. *Mon. Wea. Rev.*, **144**, 615–626, https://doi.org/10.1175/MWR-D-15-0218.1.

——, J. Nattala, and M. K. Tippett, 2014: Skill improvement from increased ensemble size and model diversity. *Geophys. Res. Lett.*, **41**, 7331–7342, https://doi.org/10.1002/2014GL060133.

Doblas-Reyes, F. J., R. Hagedorn, and T. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting—II. Calibration and combination. *Tellus*, **57A**, 234–252, https://doi.org/10.3402/tellusa.v57i3.14658.

——, and Coauthors, 2013: Initialized near-term regional climate change prediction. *Nat. Commun.*, **4**, 1715, https://doi.org/10.1038/ncomms2704.

Douville, H., A. Voldoire, and O. Geoffroy, 2015: The recent global warming hiatus: What is the role of Pacific variability? *Geophys. Res. Lett.*, **42**, 880–888, https://doi.org/10.1002/2014GL062775.

Dunstone, N., D. Smith, A. Scaife, L. Hermanson, R. Eade, N. Robinson, M. Andrews, and J. Knight, 2016: Skilful predictions of the winter North Atlantic Oscillation one year ahead. *Nat. Geosci.*, **9**, 809–814, https://doi.org/10.1038/ngeo2824.

——, and Coauthors, 2020: Skilful interannual climate prediction from two large initialised model ensembles. *Environ. Res. Lett.*, **15**, 094083, https://doi.org/10.1088/1748-9326/ab9f7d.

Eade, R., D. Smith, A. Scaife, E. Wallace, N. Dunstone, L. Hermanson, and N. Robinson, 2014: Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophys. Res. Lett.*, **41**, 5620–5628, https://doi.org/10.1002/2014GL061146.

Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, 2016: Overview of the Coupled Model Intercomparison Project phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.*, **9**, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016.

Fricker, T. E., C. A. Ferro, and D. B. Stephenson, 2013: Three recommendations for evaluating climate predictions. *Meteor. Appl.*, **20**, 246–255, https://doi.org/10.1002/met.1409.

Frumkin, H., J. Hess, G. Luber, J. Malilay, and M. McGeehin, 2008: Climate change: The public health response. *Amer. J. Public Health*, **98**, 435–445, https://doi.org/10.2105/AJPH.2007.119362.

Fyfe, J. C., and Coauthors, 2016: Making sense of the early-2000s warming slowdown. *Nat. Climate Change*, **6**, 224–228, https://doi.org/10.1038/nclimate2938.

García-Serrano, J., V. Guemas, and F. J. Doblas-Reyes, 2015: Added-value from initialization in predictions of Atlantic multi-decadal variability. *Climate Dyn.*, **44**, 2539–2555, https://doi.org/10.1007/s00382-014-2370-7.

Goddard, L., and Coauthors, 2013: A verification framework for interannual-to-decadal predictions experiments. *Climate Dyn.*, **40**, 245–272, https://doi.org/10.1007/s00382-012-1481-2.

Hagedorn, R., F. J. Doblas-Reyes, and T. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept. *Tellus*, **57A**, 219–233, https://doi.org/10.3402/tellusa.v57i3.14657.

Hazeleger, W., V. Guemas, B. Wouters, S. Corti, I. Andreu-Burillo, F. J. Doblas-Reyes, K. Wyser, and M. Caian, 2013: Multiyear climate predictions using two initialization strategies. *Geophys. Res. Lett.*, **40**, 1794–1798, https://doi.org/10.1002/grl.50355.

Hemri, S., and Coauthors, 2020: How to create an operational multi-model of seasonal forecasts? *Climate Dyn.*, **55**, 1141–1157, https://doi.org/10.1007/s00382-020-05314-2.

Kirtman, B., and A. Pirani, 2009: The state of the art of seasonal prediction: Outcomes and recommendations from the First World Climate Research Program Workshop on Seasonal Prediction. *Bull. Amer. Meteor. Soc.*, **90**, 455–458, https://doi.org/10.1175/2008BAMS2707.1.

Kobayashi, S., and Coauthors, 2015: The JRA-55 reanalysis: General specifications and basic characteristics. *J. Meteor. Soc. Japan*, **93**, 5–48, https://doi.org/10.2151/jmsj.2015-001.

Lenssen, N. J. L., G. A. Schmidt, J. E. Hansen, M. J. Menne, A. Persin, R. Ruedy, and D. Zyss, 2019: Improvements in the GISTEMP uncertainty model. *J. Geophys. Res. Atmos.*, **124**, 6307–6326, https://doi.org/10.1029/2018JD029522.

Lledó, L., I. Cionni, V. Torralba, P.-A. Bretonnière, and M. Samsó, 2020: Seasonal prediction of Euro-Atlantic teleconnections from multiple systems. *Environ. Res. Lett.*, **15**, 074009, https://doi.org/10.1088/1748-9326/ab87d2.

Mahmood, R., M. G. Donat, P. Ortega, F. J. Doblas-Reyes, and Y. Ruprich-Robert, 2021: Constraining decadal variability yields skillful projections of near-term climate change. *Geophys. Res. Lett.*, **48**, e2021GL094915, https://doi.org/10.1029/2021GL094915.

Manzanas, R., and J. M. Gutiérrez, 2018: Process-conditioned bias correction for seasonal forecasting: A case-study with ENSO in Peru. *Climate Dyn.*, **52**, 1673–1683, https://doi.org/10.1007/s00382-018-4226-z.

Marcos, R., M. C. Llasat, P. Quintana-Seguí, and M. Turco, 2018: Use of bias correction techniques to improve seasonal forecasts for reservoirs—A case-study in northwestern Mediterranean. *Sci. Total Environ.*, **610–611**, 64–74, https://doi.org/10.1016/j.scitotenv.2017.08.010.

Mauritsen, T., and Coauthors, 2019: Developments in the MPI-M Earth system model version 1.2 (MPI-ESM1.2) and its response to increasing CO_2. *J. Adv. Model. Earth Syst.*, **11**, 998–1038, https://doi.org/10.1029/2018MS001400.

Meehl, G. A., and Coauthors, 2014: Decadal climate prediction: An update from the trenches. *Bull. Amer. Meteor. Soc.*, **95**, 243–267, https://doi.org/10.1175/BAMS-D-12-00241.1.

Menne, M. J., C. N. Williams, B. E. Gleason, J. J. Rennie, and J. H. Lawrimore, 2018: The Global Historical Climatology Network monthly temperature dataset, version 4. *J. Climate*, **31**, 9835–9854, https://doi.org/10.1175/JCLI-D-18-0094.1.

Mishra, N., C. Prodhomme, and V. Guemas, 2018: Multi-model skill assessment of seasonal temperature and precipitation forecasts over Europe. *Climate Dyn.*, **52**, 4207–4225, https://doi.org/10.1007/s00382-018-4404-z.

Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones, 2012: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *J. Geophys. Res.*, **117**, D08101, https://doi.org/10.1029/2011JD017187.

Müller, W. A., and Coauthors, 2018: A higher-resolution version of the Max Planck Institute Earth System Model (MPI-ESM1.2-HR). *J. Adv. Model. Earth Syst.*, **10**, 1383–1413, https://doi.org/10.1029/2017MS001217.

Murphy, A. H., 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon. Wea. Rev.*, **116**, 2417–2424, https://doi.org/10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2.

——, 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293, https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

Papalexiou, S. M., C. R. Rajulapati, M. P. Clark, and F. Lehner, 2020: Robustness of CMIP6 historical global mean temperature simulations: Trends, long-term persistence, autocorrelation, and distributional shape. *Earth's Future*, **8**, e2020EF001667, https://doi.org/10.1029/2020EF001667.

Pasternack, A., J. Bhend, M. A. Liniger, H. W. Rust, W. A. Müller, and U. Ulbrich, 2018: Parametric decadal climate forecast recalibration (DeFoReSt 1.0). *Geosci. Model Dev.*, **11**, 351–368, https://doi.org/10.5194/gmd-11-351-2018.

Paxian, A., and Coauthors, 2019: User-oriented global predictions of the GPCC drought index for the next decade. *Meteor. Z.*, **28**, 3–21, https://doi.org/10.1127/metz/2018/0912.

Pérez-Zanón, N., and Coauthors, 2021: The CSTools (v4.0) toolbox: From climate forecasts to climate forecast information. *Geosci. Model Dev. Discuss.*, https://doi.org/10.5194/gmd-2021-368.

Polkova, I., and Coauthors, 2019: Initialization and ensemble generation for decadal climate predictions: A comparison of different methods. *J. Adv. Model. Earth Syst.*, **11**, 149–172, https://doi.org/10.1029/2018MS001439.

Power, S., T. Casey, C. Folland, A. Colman, and V. Mehta, 1999: Inter-decadal modulation of the impact of ENSO on Australia. *Climate Dyn.*, **15**, 319–324, https://doi.org/10.1007/s003820050284.

Scaife, A. A., and D. Smith, 2018: A signal-to-noise paradox in climate science. *npj Climate Atmos. Sci.*, **1**, 28, https://doi.org/10.1038/s41612-018-0038-4.

Schneider, U., A. Becker, P. Finger, A. Meyer-Christoffer, and M. Ziese, 2018: GPCC full data monthly product version 2018 at 1.0°: Monthly land-surface precipitation from rain-gauges built on GTS-based and historical data. Global Precipitation Climatology Centre, accessed 6 April 2020, https://doi.org/10.5676/DWD_GPCC/FD_M_V2018_100.

Sellar, A. A., and Coauthors, 2020: Implementation of U.K. Earth system models for CMIP6. *J. Adv. Model. Earth Syst.*, **12**, e2019MS001946, https://doi.org/10.1029/2019MS001946.

Sgubin, G., D. Swingedouw, L. F. Borchert, M. B. Menary, T. Noël, H. Loukos, and J. Mignot, 2021: Systematic investigation of skill opportunities in decadal prediction of air temperature over Europe. *Climate Dyn.*, **57**, 3245–3263, https://doi.org/10.1007/s00382-021-05863-0.

Sheen, K. L., D. M. Smith, N. J. Dunstone, R. Eade, D. P. Rowell, and M. Vellinga, 2017: Skilful prediction of Sahel summer rainfall on inter-annual and multi-year timescales. *Nat. Commun.*, **8**, 14966, https://doi.org/10.1038/ncomms14966.

Si, D., A. Hu, H. Wang, and Q. Chao, 2019: Predicting the Atlantic multidecadal variability from initialized simulations. *J. Climate*, **32**, 8701–8711, https://doi.org/10.1175/JCLI-D-19-0055.1.

Slingo, J., and T. Palmer, 2011: Uncertainty in weather and climate prediction. *Philos. Trans. Roy. Soc. London*, **A369**, 4751–4767, https://doi.org/10.1098/rsta.2011.0161.

Smith, D. M., R. Eade, and H. Pohlmann, 2013: A comparison of full-field and anomaly initialization for seasonal to decadal climate prediction. *Climate Dyn.*, **41**, 3325–3338, https://doi.org/10.1007/s00382-013-1683-2.

——, and Coauthors, 2019: Robust skill of decadal climate predictions. *npj Climate Atmos. Sci.*, **2**, 13, https://doi.org/10.1038/s41612-019-0071-y.

——, and Coauthors, 2020: North Atlantic climate far more predictable than models imply. *Nature*, **583**, 796–800, https://doi.org/10.1038/s41586-020-2525-0.

Solaraju-Murali, B., L.-P. Caron, N. Gonzalez-Reviriego, and F. J. Doblas-Reyes, 2019: Multi-year prediction of European summer drought conditions for the agricultural sector. *Environ. Res. Lett.*, **14**, 124014, https://doi.org/10.1088/1748-9326/ab5043.

——, and Coauthors, 2021: Multi-annual prediction of drought and heat stress to support decision making in the wheat sector. *npj Climate Atmos. Sci.*, **4**, 34, https://doi.org/10.1038/s41612-021-00189-4.

Sospedra-Alfonso, R., and G. J. Boer, 2020: Assessing the impact of initialization on decadal prediction skill. *Geophys. Res. Lett.*, **47**, e2019GL086361, https://doi.org/10.1029/2019GL086361.

Swart, N. C., and Coauthors, 2019: The Canadian Earth System Model version 5 (CanESM5.0.3). *Geosci. Model Dev.*, **12**, 4823–4873, https://doi.org/10.5194/gmd-12-4823-2019.

Tatebe, H., and Coauthors, 2019: Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in MIROC6. *Geosci. Model Dev.*, **12**, 2727–2765, https://doi.org/10.5194/gmd-12-2727-2019.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, **93**, 485–498, https://doi.org/10.1175/BAMS-D-11-00094.1.

Tian, T., S. Yang, M. P. Karami, F. Massonnet, T. Kruschke, and T. Koenigk, 2021: Benefits of sea ice initialization for the interannual-to-decadal climate prediction skill in the Arctic in EC-Earth3. *Geosci. Model Dev.*, **14**, 4283–4305, https://doi.org/10.5194/gmd-14-4283-2021.

Tippett, M. K., and A. G. Barnston, 2008: Skill of multi-model ENSO probability forecasts. *Mon. Wea. Rev.*, **136**, 3933–3946, https://doi.org/10.1175/2008MWR2431.1.

Tokarska, K. B., M. B. Stolpe, S. Sippel, E. M. Fischer, C. J. Smith, F. Lehner, and R. Knutti, 2020: Past warming trend constrains future warming in CMIP6 models. *Sci. Adv.*, **6**, eaaz9549, https://doi.org/10.1126/sciadv.aaz9549.

Tommasi, D., and Coauthors, 2017: Managing living marine resources in a dynamic environment: The role of seasonal to decadal climate forecasts. *Prog. Oceanogr.*, **152**, 15–49, https://doi.org/10.1016/j.pocean.2016.12.011.

Torralba, V., F. J. Doblas-Reyes, D. MacLeod, I. Christel, and M. Davis, 2017: Seasonal climate prediction: A new source of information for the management of wind energy resources. *J. Appl. Meteor. Climatol.*, **56**, 1231–1247, https://doi.org/10.1175/JAMC-D-16-0204.1.

Trenberth, K. E., and D. J. Shea, 2006: Atlantic hurricanes and natural variability in 2005. *Geophys. Res. Lett.*, **33**, L12704, https://doi.org/10.1029/2006GL026894.

van Schaeybroeck, B., and S. Vannitsem, 2011: Post-processing through linear regression. *Nonlinear Processes Geophys.*, **18**, 147–160, https://doi.org/10.5194/npg-18-147-2011.

——, and ——, 2015: Ensemble post-processing using member-by-member approaches: Theoretical aspects. *Quart. J. Roy. Meteor. Soc.*, **141**, 807–818, https://doi.org/10.1002/qj.2397.

Vera, C., and Coauthors, 2010: Needs assessment for climate information on decadal timescales and longer. *Procedia Environ. Sci.*, **1**, 275–286, https://doi.org/10.1016/j.proenv.2010.09.017.

von Storch, H., and F. W. Zwiers, 1999: *Statistical Analysis in Climate Research*. Cambridge University Press, 484 pp.

Wang, C., B. J. Soden, W. Yang, and G. A. Vecchi, 2021: Compensation between cloud feedback and aerosol–cloud interaction in CMIP6 models. *Geophys. Res. Lett.*, **48**, e2020GL091024, https://doi.org/10.1029/2020GL091024.

Wanner, H., S. Brönnimann, C. Casty, D. Gyalistras, J. Luterbacher, C. Schmutz, D. B. Stephenson, and E. Xoplaki, 2001: North Atlantic Oscillation—Concepts and studies. *Surv. Geophys.*, **22**, 321–381, https://doi.org/10.1023/A:1014217317898.

Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2009: Seasonal ensemble forecasts: Are recalibrated single models better than multi-models? *Mon. Wea. Rev.*, **137**, 1460–1479, https://doi.org/10.1175/2008MWR2773.1.

Wilks, D. S., 2011: Forecast verification. *Statistical Methods in the Atmospheric Sciences*, Academic Press, 301–394, https://doi.org/10.1016/B978-0-12-385022-5.00008-7.

Wu, T., and Coauthors, 2019: The Beijing Climate Center Climate System Model (BCC-CSM): The main progress from CMIP5 to CMIP6. *Geosci. Model Dev.*, **12**, 1573–1600, https://doi.org/10.5194/gmd-12-1573-2019.

Yeager, S. G., and Coauthors, 2018: Predicting near-term changes in the Earth system: A large ensemble of initialized decadal prediction simulations using the Community Earth System Model. *Bull. Amer. Meteor. Soc.*, **99**, 1867–1886, https://doi.org/10.1175/BAMS-D-17-0098.1.

Yukimoto, S., and Coauthors, 2019: The Meteorological Research Institute Earth System Model version 2.0, MRI-ESM2.0: Description and basic evaluation of the physical component. *J. Meteor. Soc. Japan*, **97**, 931–965, https://doi.org/10.2151/jmsj.2019-051.

Zhao, T., J. C. Bennett, Q. J. Wang, A. Schepen, A. W. Wood, D. E. Robertson, and M.-H. Ramos, 2017: How suitable is quantile mapping for postprocessing GCM precipitation forecasts? *J. Climate*, **30**, 3185–3196, https://doi.org/10.1175/JCLI-D-16-0652.1.

Chapter 4

Representation and annual to decadal predictability of Euro-Atlantic weather regimes in the CMIP6 version of the EC-Earth coupled climate model

This chapter has been published as peer-reviewed article as:

Delgado-Torres, C., Verfaillie, D., Mohino, E., and Donat, M. G. (2022). Representation and annual to decadal predictability of Euro-Atlantic weather regimes in the CMIP6 version of the EC-Earth coupled climate model. Journal of Geophysical Research: Atmospheres, 127, e2022JD036673. https://doi.org/10.1029/2022JD036673

The supplementary material can be found in Appendix C.

4.1. Main objectives

- Identify the four weather regimes over the Euro-Atlantic region during different seasons.

- Assess the spatial representation and climatological frequency of such weather regimes in the EC-Earth3 forecast system.
- Evaluate the ability to predict the inter-annual to decadal variability of the weather regimes' seasonal frequency of occurrence.
- Analyse the impact of model initialisation for the representation and prediction skill.

4.2. Main outcomes

- The EC-Earth3 forecast system correctly represents the spatial patterns and climatological occurrence frequencies of the four weather regimes.
- The skill in predicting the inter-annual to decadal variations of the seasonal frequencies is generally low.
- The model initialisation does not improve the prediction skill.
- The teleconnections between the weather regimes and the North Atlantic SST are generally not reproduced by the model, which might limit the prediction skill.

JGR Atmospheres

RESEARCH ARTICLE
10.1029/2022JD036673

Key Points:

- The EC-Earth3 model simulates the spatial patterns and climatological frequencies of the Euro-Atlantic weather regimes realistically
- Correlations between simulated and observed frequencies of weather regimes on inter-annual to decadal time scales are generally low
- Model initialization does not significantly alter the skill in predicting the spatial patterns and temporal variations of the weather regimes

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

C. Delgado-Torres,
carlos.delgado@bsc.es

Citation:

Delgado-Torres, C., Verfaillie, D., Mohino, E., & Donat, M. G. (2022). Representation and annual to decadal predictability of Euro-Atlantic weather regimes in the CMIP6 version of the EC-Earth coupled climate model. *Journal of Geophysical Research: Atmospheres*, *127*, e2022JD036673. https://doi.org/10.1029/2022JD036673

Received 23 FEB 2022
Accepted 5 JUL 2022

Representation and Annual to Decadal Predictability of Euro-Atlantic Weather Regimes in the CMIP6 Version of the EC-Earth Coupled Climate Model

C. Delgado-Torres[1,2], D. Verfaillie[1,3], E. Mohino[2], and M. G. Donat[1,4]

[1]Barcelona Supercomputing Center (BSC), Barcelona, Spain, [2]Department of Physics of the Earth and Astrophysics, Universidad Complutense de Madrid, Madrid, Spain, [3]Aix-Marseille University, CNRS, IRD, Collège de France, INRAE, CEREGE, Aix-en-Provence, France, [4]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

Abstract Weather regimes are large-scale atmospheric circulation states that frequently occur in the climate system with persistence and recurrence, and are associated with the occurrence of specific local weather conditions. This study evaluates the representation of the four Euro-Atlantic weather regimes in uninitialized historical forcing simulations and initialized decadal predictions performed with the EC-Earth3 coupled climate model. The four weather regimes are the positive and negative phases of the North Atlantic Oscillation (NAO+ and NAO−, respectively), Blocking, and Atlantic Ridge in winter; and the NAO−, Blocking, Atlantic Ridge, and Atlantic Low in summer. We also analyze the impact that the model initialization toward the observed state of the climate system has on the ability to predict the variability of the weather regimes' seasonal frequency of occurrence. We find that the EC-Earth3 model correctly reproduces the spatial patterns and climatological occurrence frequencies of the four weather regimes. By contrast, the skill in predicting the inter-annual to decadal variations of the weather regimes' seasonal frequencies is generally low, and the initialization does not significantly improve such skill. The observed teleconnections between the weather regimes and the North Atlantic sea surface temperatures are generally not reproduced by the model, which could be a reason for the low skill in predicting the temporal variations of the weather regime frequencies.

1. Introduction

Weather regimes are a set of climate states that occur more frequently due to either more persistence or more recurrence than other possible states of the climate system (Christensen et al., 2013; Cortesi et al., 2019; Michelangeli et al., 1995). They provide a simplified description of the atmospheric flow variability and describe large-scale conditions that can be associated with local weather. Skillful climate predictions of the weather regimes could thus be used as a source of predictability for local climate conditions (Hertig & Jacobeit, 2014). Therefore, a correct representation and prediction of weather regimes by climate models could be translated into useful climate information for decision-makers. At inter-annual to decadal time scales, such climate information could be provided by decadal climate predictions, which aim at filling the gap between seasonal predictions and climate projections (Kirtman et al., 2013).

The atmospheric states gather naturally into four well-defined, statistically robust clusters over the Euro-Atlantic sector, with a typical persistence of 3–7 days (Michelangeli et al., 1995). The four weather regimes are different across the seasons. In winter, the positive and negative phases of the North Atlantic Oscillation (NAO+ and NAO−, respectively), the Blocking, and the Atlantic Ridge are identified in the literature (Cattiaux, Quesada, et al., 2013; Cortesi et al., 2019, 2021; Dawson et al., 2012; Ferranti et al., 2015). In summer, the Euro-Atlantic weather regimes are the Atlantic Low, the NAO−, the Blocking, and the Atlantic Ridge (Cassou et al., 2005; Cattiaux, Quesada, et al., 2013). However, the existence and the optimal number of regimes is a subject with conflicting results (see, e.g., Christiansen [2007] for a review), and some studies even reject the existence of weather regimes (Stephenson et al., 2004). Still, a classification of flow regimes can be useful to characterize large-scale weather conditions and to evaluate how climate models represent large-scale atmospheric flow.

The simulation and prediction of the Euro-Atlantic weather regimes by climate models are hindered by the characteristic biases that models tend to show in the representation of atmospheric flow over the Euro-Atlantic domain (Walz et al., 2018). In particular, evaluations of climate models have often found an underestimation of the occurrence frequency and persistence of blocking events (Cattiaux, Quesada, et al., 2013; D'Andrea et al., 1998; Masato et al., 2013; Schiemann et al., 2017), related to an overestimation of zonal (westerly) flow regimes in Europe (Donat et al., 2010). This paper therefore provides a systematic evaluation of climatological characteristics of weather regimes in the latest version of the EC-Earth model.

Climate predictions have been shown to skillfully predict essential climate variables at sub-seasonal to seasonal (Mariotti et al., 2020) and inter-annual to decadal (Kushnir et al., 2019) time scales, which can be useful to underpin decision-making in a wide range of sectors (Merryfield et al., 2020). These climate predictions use, in addition to external forcings (which provide skill in climate projections), predictability provided by slow variations of different components of the climate system, especially the ocean, land surface, and sea ice (Meehl et al., 2009, 2021). In order to make use of these potential sources of predictability, climate models are used to compute the future evolution of the climate system by integrating them forward in time from a set of observation-based initial conditions, which is referred to as model initialization.

On the seasonal time scale, Cortesi et al. (2017) showed skill in reproducing the weather regimes' spatial patterns, climatological frequencies, persistences, and transitions probabilities during winter, spring, and summer with the ECMWF seasonal forecasting system S4 (Molteni et al., 2011). However, they found low skill in reproducing the monthly frequency variations of the weather regimes. Carvalho-Oliveira et al. (2022) assessed the representation of the summer spatial patterns of the weather regimes within the MPI-ESM-MR seasonal forecasting system (Giorgetta et al., 2013), finding that the Atlantic Ridge regime shows the highest agreement with the reanalysis, while the Atlantic Low regime shows the lowest agreement. On inter-annual to decadal time scales, while decadal predictions in particular of temperature are skillful in many regions due to external forcings, model initialization has also been shown to add prediction skill for several climate variables (Smith et al., 2019). Also, some selected characteristics of large-scale atmospheric flow have been found to be predictable. For instance, Athanasiadis et al. (2020) have shown decadal prediction skill for the High Latitude Blocking and the NAO index during winter. Significant skill in predicting the NAO index was also found by Smith et al. (2020) using a very large ensemble based on multiple models and applying post-processing techniques to overcome the signal-to-noise problem in climate models (Scaife & Smith, 2018). However, a systematic evaluation addressing the decadal prediction skill of the objectively identified Euro-Atlantic weather regimes in different seasons as well as the impact that the initialization has on the skill is, to our knowledge, still missing.

The aim of this paper is twofold. On the one hand, we evaluate the fidelity of the EC-Earth3 model in reproducing the climatological patterns and frequencies of the Euro-Atlantic weather regimes. On the other hand, we compare the weather regimes in initialized decadal predictions and transient historical forcing simulations to quantify the impact of the model initialization on the skill in predicting the variability of the weather regimes on inter-annual to decadal time scales.

2. Data

In this study, a ten-member ensemble of initialized decadal predictions (DCPP; Boer et al., 2016) and a ten-member ensemble of non-initialized historical simulations (HIST) performed with the version 3.3 of the EC-Earth model (Döscher et al., 2022) in the framework of the Coupled Model Intercomparison Project Phase 6 (CMIP6; Eyring et al., 2016) are used. The historical simulations are started from different states of a pre-industrial control run with forcing held fix to 1850 levels, and are forced by observed external forcings until 2014. The decadal predictions have been produced by initializing the EC-Earth3 model every year from 1960 in November using a full-field initialization technique, and provide predictions for the next 11 years (Bilbao et al., 2021).

Two different reanalyses are used as reference datasets to evaluate the skill of the decadal predictions and historical simulations in representing the spatial patterns and climatological frequencies, and in predicting the variations of the seasonal frequency of occurrence of the weather regimes. These reanalyses are the Japanese 55-year Reanalysis (JRA-55; Kobayashi et al., 2015) and the NCEP/NCAR Reanalysis 1 (NCEP1; Kalnay et al., 1996). The main reason for choosing the JRA-55 and the NCEP1 reanalyses is their temporal coverage matching that of the decadal predictions. The JRA-55 data set is often preferred over others like NCEP1 because of its higher spatial resolution, making this reanalysis suitable for weather regime classification (Cortesi et al., 2019). However, here we use both reanalyses to take into account the observational uncertainty and assess the robustness of the results to the choice of the reference data set. Stryhal and Huth (2017) made a comparison between five different reanalyses (in which both the JRA-55 and NCEP1 were included) for the classification of patterns in the Euro-Atlantic area in winter. They found that there is very little difference between the daily pattern classification, especially in mid-latitudes of the Northern Hemisphere (less than 8% of days classified differently), where there is a dense coverage of high-quality observations.

3. Methods

We use daily fields of sea level pressure (from the reanalyses and EC-Earth3 simulations) to compute the weather regimes over the Euro-Atlantic region delimited between 27°–81°N and 85.5°W–45°E, as in Cortesi et al. (2019). Some studies used the geopotential height at 500 or 700 hPa instead to compute the weather regimes (Casola & Wallace, 2007; Kageyama et al., 1999). However, we use the sea level pressure because it is a variable less affected by global warming than the geopotential height (Cortesi et al., 2021; Torralba et al., 2021) and the clustering will thus be more accurate over the long period considered. The winter (defined as December-January-February; DJF) and summer (defined as June-July-August; JJA) weather regimes are obtained during the 1965–2014 period (50 years) for the reanalyses, the historical simulations and the decadal predictions. For the reanalyses and the historical simulations, seasonal averages and multi-annual averages of seasonal means are considered. In the case of the decadal predictions, the first five individual forecast winters and summers, and all their possible multi-year averages have been included in the assessment. In addition, the analysis also includes the winters and summers for the average of forecast years 1–10 to account for the full predicted decade. Since different forecast periods are assessed, different start dates have been selected in each case in order to always evaluate the predictions over the same calendar period (i.e., over the 1965–2014 period). The evaluation period starts in 1965 because forecast year 5 is the last forecast year evaluated individually (and 1965 is the forecast year 5 from the first initialization, i.e., start date November 1960), and the evaluation period ends in 2014 because the historical simulations are run until that year.

The simulated data have been interpolated onto the spatial grid of the reanalyses to obtain the spatial patterns of the weather regimes in the same grid resolution and evaluate their spatial representation. Also, the daily pressure fields have been converted into daily anomalies by subtracting the daily climatology. The daily climatology has been computed over the whole period and smoothed by applying a Locally Estimated Scatterplot Smoothing filter (LOESS; Cleveland & Devlin, 1988) with a smoothing span of 1 to remove the short-term variability and retain the seasonal cycle (Torralba, 2019).

In order to classify the daily anomaly maps and compute the weather regimes, the *k*-means clustering algorithm (Michelangeli et al., 1995; Philipp et al., 2010) has been applied independently for each season and, in case of the historical simulations and decadal predictions, also independently for each ensemble member. The *k*-means algorithm aims at arranging a set of daily maps within groups, called clusters, seeking the most steady states. The algorithm minimizes the sum of the squared distances from each map to the centroid of the clusters to which they belong, providing the common spatial patterns in the analyzed area. The anomalies are previously weighted by the cosine of the latitude to take into account the different sizes of the grid boxes in the region considered (Cortesi et al., 2019, 2021; Falkena et al., 2020).

The number of clusters k to generate has to be specified in advance. Fereday et al. (2008) assessed the optimal number of clusters and concluded that there is no objective choice for the number of clusters. For small k values, the full range of patterns is not correctly represented, while if a large k is chosen different clusters can look very similar. The number of clusters generated in this study is $k = 4$, typically used in the literature to assess the Euro-Atlantic weather regimes (Cassou et al., 2005; Cattiaux, Quesada, et al., 2013; Cortesi et al., 2019, 2021; Dawson et al., 2012; Ferranti et al., 2015; Hertig & Jacobeit, 2014). In addition, to confirm that the four weather regimes are obtained even if the *k*-means algorithm is set to define more clusters, the same procedure has been used but applying the *k*-means algorithm to identify five clusters ($k = 5$). Similarly, the first and second halves of the evaluation period (i.e., 1965–1989 and 1990–2014) have been used to obtain the four weather regimes ($k = 4$, as in the rest of the study) to confirm that such weather regimes are present during the whole period.

The observed and simulated weather regimes are independently computed by cluster analysis with the *k*-means algorithm applied to the daily anomalies fields to obtain the spatial patterns. Then, the daily maps are projected onto the four patterns using the minimum Euclidean distance method. Both minimal-correlation and minimal-persistence filters are applied to both observed and simulated projected daily maps to filter out the days that do not actually belong to any of the clusters. The minimal-correlation filter consists of de-classifying those days with a spatial correlation (measured with the spatial Anomaly Correlation Coefficient, ACC; Wilks, 2011) lower than 0.25 with the cluster it was assigned. The minimal-persistence filter de-classifies those days that do not belong to a spell of at least 3 days. This methodology is similar to that used in Cattiaux, Douville, and Peings (2013) but without previously applying Empirical Orthogonal Functions to take into account the extreme sea level pressure values. This method tests whether the EC-Earth3 model correctly represents the weather regimes' spatial patterns by itself. Additionally, the same methodology has been employed but projecting the simulated daily maps onto the observed clusters (i.e., only applying the *k*-means algorithm to the reanalysis data) to assess how the simulated daily maps fit onto the observed patterns. Results in the main text correspond to the first approach, that is, projecting the simulated and observed fields onto the simulated and observed patterns, respectively. The results for the alternative method, that is, projecting the simulated fields onto the observed patterns, are included in Supporting Information S1.

Once each day is assigned to one of the four seasonal clusters (or to the unclassified cluster), the spatial patterns of the weather regimes are computed as the composites (averaged daily maps) of all the days that have been assigned to a cluster. The relative seasonal frequency of occurrence of each weather regime is computed as the percentage of the number of days per season assigned to each cluster. For the simulations, since the clustering has been applied to each ensemble member separately, the frequency time series and the spatial patterns are calculated as the ensemble mean.

The spatial correlation between the simulated and observed patterns is estimated with the spatial ACC. This metric has been computed by accounting for the different sizes of the grid cells. The significance of the spatial correlation has been assessed by estimating the confidence intervals at the 95% confidence level computed by a Fisher transformation. If the confidence interval includes zero, the spatial correlation is not significant. The spatial correlation difference between the historical simulations and decadal predictions is significant if their confidence intervals do not overlap. A two-sided Kolmogorov-Smirnov test has been applied to assess whether the simulated and observed distributions of the seasonal frequencies could have been sampled from the same continuous distribution at the 95% confidence level. In addition, a two-sided *t*-test has been used to assess whether the mean values of the simulated and observed distributions are statistically different at the 95% confidence level. In order to estimate the temporal correlation, the Pearson's correlation coefficient is used. A two-tailed *t*-test is applied to analyze whether the temporal correlation between experiments and reanalyses is statistically significant at the 95% confidence level. The Fisher's *z* transformed correlations divided by the standard error of the difference (Wilks, 2011) are used to assess whether the correlations obtained with the historical simulations and the decadal predictions are statistically different at the 95% confidence level. In order to avoid a potential overestimation of the statistical significance due to spatial patterns and time series autocorrelation, the effective degrees of freedom have been used for the significance tests. Such effective degrees of freedom have been calculated over the reanalysis data following Zwiers and von Storch (1995). In the case of the spatial patterns, the methodology has been applied over the vector created by concatenating all latitudes one after the other. It should be noted that part of the information related to different latitudes might be lost when computing the effective number of degrees of freedom for the spatial patterns.

4. Results

4.1. Evaluation of the Spatial Patterns and Climatological Frequencies of the Weather Regimes

We first evaluate two climatological aspects of weather regimes: the spatial patterns (i.e., the composites pressure maps of all the days that belong to each weather regime) and the climatological seasonal frequencies of occurrence.

The spatial patterns of the observed weather regimes during the winter and summer seasons identified in the JRA-55 reanalysis are shown in Figure 1. The four clusters obtained for each season correspond to the NAO+, NAO−, Blocking, and Atlantic Ridge during winter; and the Atlantic Low, NAO−, Blocking, and Atlantic Ridge during summer. The regime patterns show that the amplitude of the clusters' anomalies is higher during the winter season, when they are more persistent in time and have a stronger effect on local climate (Ferranti et al., 2015; Fil & Dubus, 2005). Compared with the spatial patterns using the NCEP1 reanalysis (Figure S1 in Supporting Information S1), there is a high consistency for both seasons, indicating the robustness of the identified weather regimes to the choice of the reference data set.

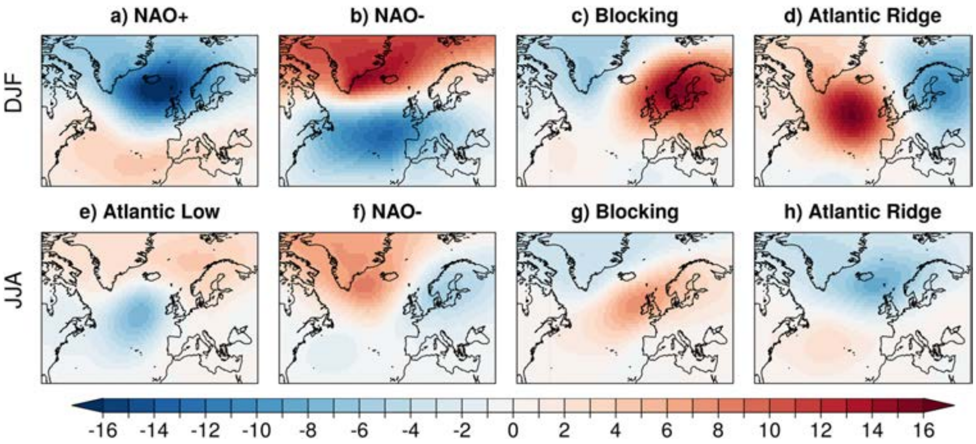


Figure 1. Spatial patterns of the observed Euro-Atlantic weather regimes (computed as the averaged sea level pressure anomalies, in hPa, of all the days classified onto each cluster) obtained with the JRA-55 reanalysis for winter and summer during the 1965–2014 period.

We test the sensitivity of the clustering algorithm results to the number of identified clusters and the period used to identify the clusters. We find that the same four observed weather regimes for winter are obtained when asking for five clusters (Figure S2 in Supporting Information S1). The additional cluster is represented as a center of anomalously negative pressure over the central part of the region considered. The sum of NAO+ and the additional cluster resembles the original NAO+ obtained with $k = 4$. When applying the k-means algorithm to the first and second halves of the evaluation period, both sets of clusters represent the same weather regimes as when using the whole period. However, some discrepancies are found in the first half (1965–1989) for NAO+ (for which the northern part of Europe contains a positive anomaly, while the anomalies over that region are slightly negative when using the whole period) and Blocking (which is displaced westward with respect to the original one). For summer (Figure S3 in Supporting Information S1), the patterns obtained with $k = 5$ and with different period halves are similar to the original ones.

The spatial patterns of the four weather regimes obtained during the whole period with both DCPP and HIST show a high agreement with the observed patterns in winter, being significant for all the forecast years and multi-year averages (Figure 2a). For summer, significant spatial correlations are also found for the NAO−, Blocking, and Atlantic Ridge regimes for both DCPP and HIST. However, the spatial correlations for DCPP are systematically high and significant for the Atlantic Low regime, except for the forecast years 1–2 (this exception is not found when using the NCEP1 reanalysis as the reference data set; Figure S4 in Supporting Information S1), while low and non-significant correlations are found for HIST. Furthermore, the spatial correlations for the different forecast years in the decadal predictions are similar for both seasons, indicating that the simulation of the weather regimes' patterns is not strongly affected by the climate drift in the decadal predictions. Similar results are found when using the NCEP1 reanalysis as the reference data set (Figure S4 in Supporting Information S1).

The ACC differences between DCPP and HIST are computed to assess whether the model initialization impacts the representation of the simulated weather regimes' patterns. The pattern correlation for the forecast periods that include more than one season is compared to the pattern correlation obtained with the historical simulations averaged over the same period length (e.g., the forecast years 1–3 are compared to 3-year averages from the historical simulations). The ACC differences (Figure 2b) indicate that DCPP and HIST perform equally in representing the spatial patterns of all weather regimes during winter, and the spatial patterns of the NAO−, Blocking and Atlantic Ridge regimes during summer (low ACC differences for all the cases, and only a few cases show significance). For the summer Atlantic Low regime, the results show an added value of model initialization in representing the spatial pattern of this weather regime. It should be noted that, even using the effective degrees of freedom to account for the auto-correlation, the sample size is still large, which means that relatively low ACC values and ACC differences may be significant.

The previous analysis (Figure 2) is focused on evaluating the agreement between the model-specific and observed spatial patterns. If, instead, we compare the observed patterns with the patterns obtained by projecting the

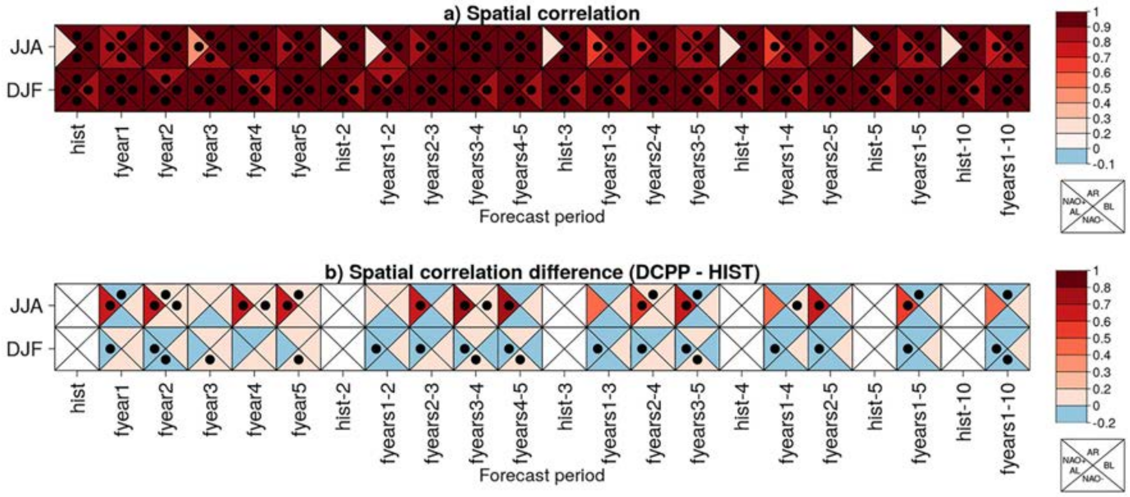


Figure 2. (a) Spatial correlation between the observed and simulated Euro-Atlantic weather regimes' patterns. (b) Difference between the ACC obtained with the decadal predictions and that obtained with the historical simulations. The evaluations period is 1965–2014. The reference data set is the JRA-55 reanalysis. The rows correspond to the winter and summer seasons. The different columns display the results for individual and multi-year averages, where hist-X corresponds to X-years averaged historical simulations and fyearsY-Z corresponds to decadal predictions for the forecast years Y-Z. Dots indicate that the correlation (a) or the correlation difference (b) is statistically significant at the 95% confidence level. The auto-correlation of the spatial patterns has been taken into account to determine the effective sample size when computing the statistical significance.

simulated daily maps onto the observed clusters (i.e., forcing the simulated daily maps to fit into the observed patterns), high and significant spatial correlations are obtained for all the weather regimes by definition. In addition, no differences between DCPP and HIST are found (Figure S5 in Supporting Information S1).

In the following, we evaluate the representation of the climatological frequency of occurrence. Figure 3 shows the distribution of the weather regimes' frequencies for the JRA-55 reanalysis, historical simulations, and decadal predictions for the first forecast year. The observed frequencies in winter show that the most frequent weather regime is NAO+ (with 23.3% of the days assigned to this cluster), followed by Blocking (20.6%), NAO− (19.1%), and Atlantic Ridge (18.1%). In summer, the most frequently observed weather regime is Blocking (21.2% of the days), followed by Atlantic Ridge (18.8%), Atlantic Low (17.8%), and NAO− (17.2%). Note that the weather regimes' frequencies do not add up to 100% due to the use of an unclassified cluster (see Section 3).

Both DCPP and HIST show a correct representation of the mean frequencies, which are found to not be significantly different to the observed mean frequencies by applying a *t*-test at the 95% confidence level. In addition, we

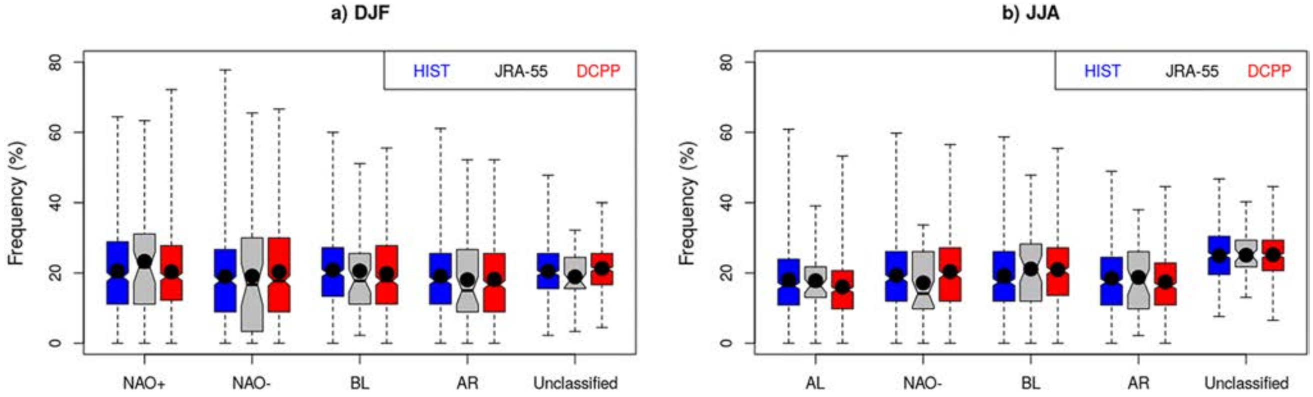


Figure 3. Box-and-whisker diagrams of the seasonal frequencies of occurrence for the Euro-Atlantic weather regimes and the unclassified cluster for winter (a) and summer (b) during the 1965–2014 period. The results are shown for the historical simulations (in blue), the JRA-55 reanalysis (in gray), and the decadal predictions (in red). Dots show the mean frequencies, while the whiskers indicate the data range between the minimum and maximum values. Model simulation boxes have been obtained using all ensemble members without averaging.

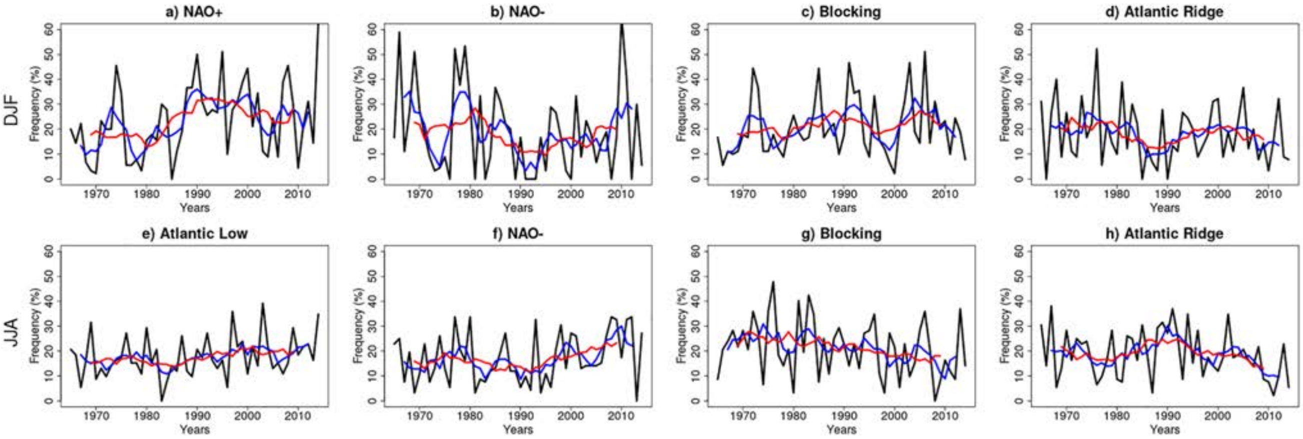


Figure 4. Time series of the weather regime frequencies of occurrence obtained with the JRA-55 reanalysis during winter (top row) and summer (bottom row) during the 1965–2014 period. In addition to 1-season averages (in black), also running averages over 5 (in blue) and 10 (in red) seasonal data points are shown.

assessed whether the simulated distributions are drawn from the same distribution as the observed ones with the Kolmogorov-Smirnov test. We find that the observed and simulated frequencies could have been drawn from the same continuous distribution, pointing to the correct representation of the frequency distribution by both DCPP and HIST. The same results are found when using the NCEP1 reanalysis as the reference data set (Figure S6 in Supporting Information S1).

The percentage of days that have been unclassified (due to either low spatial correlation or low persistence) is also shown in Figure 3. The results show that there are more unclassified days in summer than in winter, also found by Cattiaux, Douville, and Peings (2013). This happens for both the reanalyses and the climate model simulations. In the JRA-55 reanalysis, 18.9% of the days are unclassified in winter, while 25.1% of the days are unclassified during summer.

4.2. Annual to Decadal Prediction Skill of Weather Regime Frequencies

This section evaluates the skill in predicting the temporal variations of the weather regimes' seasonal occurrence frequencies (i.e., the percentage of the number of days assigned to each cluster). A skillful prediction of such variations may be useful for providing climate services based on weather regimes, which could potentially be used for downscaling the model output and as local climate predictors.

The time series obtained with the JRA-55 reanalysis show an inter-annual variability of the seasonal occurrence frequency of the weather regimes (i.e., the percentage of number of days assigned to each cluster) during both summer and winter (Figure 4). The 5-season and 10-season averages also show a temporal variability of the weather regimes' frequencies at multi-annual to decadal time scales. The time series obtained with the NCEP1 reanalysis show a similar variability as those obtained with JRA-55 (Figure S7 in Supporting Information S1). The time series show higher variability in winter than in summer at inter-annual to decadal time scales, particularly for the NAO+ and NAO− regimes. The NAO+ regime was found most frequently in the 1990s, when the NAO− was less frequent than in the previous and following decades, consistent with the time series of the NAO index shown, for example, by Athanasiadis et al. (2020) and Smith et al. (2020).

For prediction purposes, the phasing of the simulated variations of the weather regimes are compared against observations through time series correlation analysis. Figure 5a summarizes the correlation coefficients between the simulated (with both decadal predictions and historical simulations) and observed time series of the weather regimes.

In general, low correlation coefficients are found for both DCPP and HIST, with only a few cases being statistically significant. The temporal correlations for HIST mostly show coefficients close to zero. However, significantly positive correlations are found during winter for 2 and 3-year averages for the NAO− regime, and five and 10-year averages for the Atlantic Ridge regime. The correlation coefficients obtained for DCPP are generally

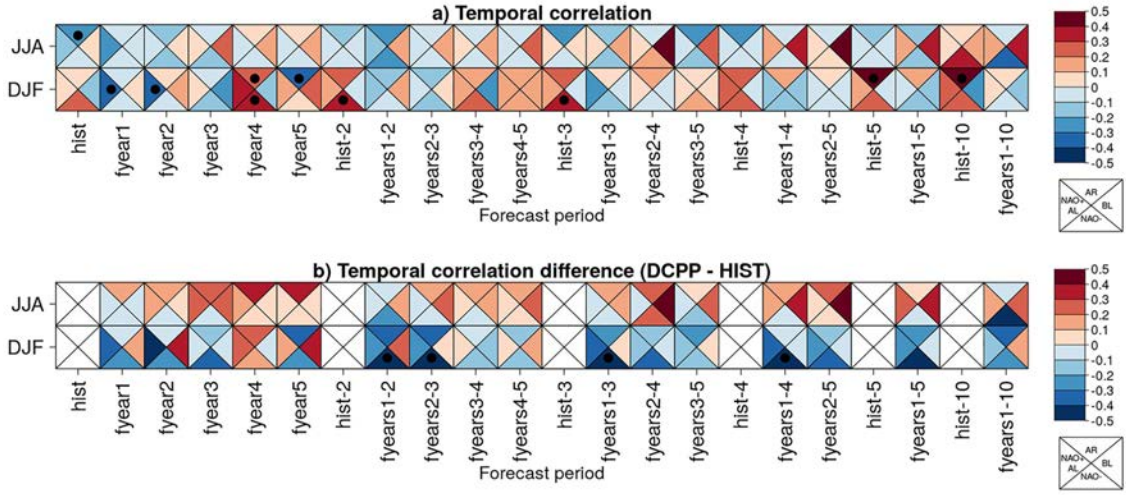


Figure 5. (a) Temporal correlation between the observed and simulated (for HIST and DCPP) Euro-Atlantic weather regimes' occurrence frequencies. (b) Difference between the correlation coefficient obtained with the decadal predictions and that obtained with the historical simulations. The evaluation period is 1965–2014. The rows correspond to the winter and summer seasons. The different columns display the results for individual and multi-year averages, where hist-X corresponds to X-years averaged historical simulations and fyearsY-Z corresponds to decadal predictions for the forecast years Y-Z. Dots indicate that the correlation (a) or the correlation difference (b) is statistically significant at the 95% confidence level. The auto-correlation of the time series has been taken into account to determine the effective sample size for the significance test.

low in winter and are significantly positive only for a few marginal cases (e.g., for NAO− and Atlantic Ridge for forecast year 4). For summer, the Blocking regime shows the highest correlation coefficients, especially when averaging at least four forecast years, but no significance is found.

In order to identify whether the model initialization has an impact on the predictability of the variability of weather regimes, the correlation differences between the decadal predictions and the historical simulations are calculated and displayed in Figure 5b. The figure does not show a clear pattern of the impact of initialization, and no significant improvements are found. In fact, the results show that the decadal predictions are less skillful in predicting the weather regimes' time series for some cases, especially for some multi-year averages of the NAO− regime during winter. The weather regime that tends to show some benefit from initialization (indicated by systematically positive correlation differences across different forecast times, although without significance) is the Blocking regime during summer. The Atlantic Ridge regime also shows an improvement from initialization (systematic for different forecast times, although not significant), but the temporal correlation is still negative for DCPP. Similar results are found when using NCEP1 as the reference data set (Figure S8 in Supporting Information S1).

4.3. Teleconnections Between the Weather Regimes and the North Atlantic SST

The slow variations of the SST provide predictability at decadal time scales (Doblas-Reyes et al., 2013; Rodwell et al., 1999; Sutton & Allen, 1997). Thus, the skill in predicting the North Atlantic SST may be transferred to some skill in predicting atmospheric circulation, as represented for example, by the Euro-Atlantic weather regimes. For this, it is needed that the model correctly reproduces (a) the observed variations of the North Atlantic SST and (b) the observed teleconnections between the weather regimes' frequencies and the North Atlantic SST.

For (a), Bilbao et al. (2021) performed a comprehensive assessment of the skill of the decadal predictions performed with the EC-Earth3 model. They showed that the model is skillful in predicting the SST over much of the North Atlantic but skill is poor in the central Subpolar North Atlantic region, and that the initialization leads to a decrease of the skill due to initialization shocks and the drift of the predictions. They discussed that these initialization issues over the North Atlantic may be improved by model development and by investigating better initialization strategies.

For (b), the observed and simulated correlations between the North Atlantic SST and the weather regime frequencies are analyzed (Figure 6). The observed teleconnection maps show that the frequencies of the weather regimes

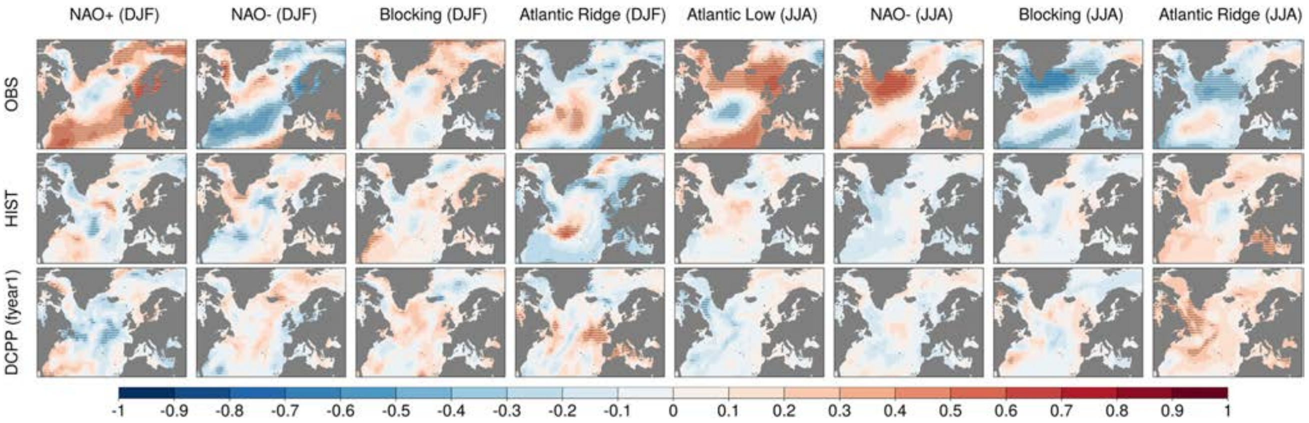


Figure 6. Correlation between the seasonal weather regimes frequency and the seasonal sea surface temperature for the JRA-55 reanalysis (top row), historical simulations (center row), and the first forecast year from decadal predictions (bottom row). In the case of the historical simulations and decadal predictions, the correlations have been computed with the ensemble mean. Dots indicate that the correlation is statistically significant at the 95% confidence level using a two-sided *t*-test. The time series' auto-correlation has been taken into account to determine the effective sample size for the significance test.

are significantly correlated with the SST over large areas of the North Atlantic Ocean, except for the winter Blocking regime which does not show significant correlations. However, this relationship is not correctly reproduced by the model, as the correlation maps show different patterns for both historical simulations and decadal predictions in comparison to that observed. Also, these teleconnections are weaker (in the sense that lower correlation coefficients are found) and in most cases not significant.

This could suggest that either (a) the weather regime frequency variations do not represent a predictable signal in relation to the SST and thus models have no possibility to predict it; (b) the relationship between the SST and weather regimes' frequency could be non-causal (SST anomalies do not force the weather regimes' frequency) and thus the phenomena phasing both are inherently unpredictable; (c) though inherently predictable, the low skill of EC-Earth3 in predicting the Subpolar North Atlantic SST (Bilbao et al., 2021) restricts the potential transferability of skill that may otherwise exist; or (d) the teleconnections are fundamentally predictable, but their signals are underestimated by the model and small compared to the noise in the ensemble.

5. Summary, Discussion, and Conclusions

This study evaluates the simulated weather regimes in the Euro-Atlantic sector in the EC-Earth3 coupled climate model in comparison to reanalyses data for the winter and summer seasons. The evaluation has been performed for both decadal predictions and historical simulations, and they are compared to assess the impact that the model initialization has on the skill.

We find that the EC-Earth3 model reproduces the spatial patterns of the four Euro-Atlantic weather regimes with high similarity to the observed patterns derived from the reanalyses. High spatial correlations are obtained with both decadal predictions and historical simulations except for the summer Atlantic Ridge regime, which shows spatial correlations close to zero for the historical simulations. Also, these spatial patterns are robust across different forecast years, indicating that they are not affected by the model drift in the decadal predictions.

Both the climatological mean frequencies and ranges of inter-annual variability in the occurrence frequencies of these weather regimes are well reproduced by the model. The number of unclassified days, due to either low spatial correlation with the observed pattern or low persistence of the regime event, is higher in summer (25.1%) than in winter (18.9%), which may be due to less intense pressure gradients during the summer season (Cattiaux, Douville, & Peings, 2013).

Regarding the skill in predicting the inter-annual and multi-annual variations of the occurrence frequencies of the weather regimes, low correlation coefficients are generally found to be, for the most part, not statistically significant. Exceptions are the winter NAO− and Atlantic Ridge regimes, which are found to be reproduced by the historical simulations with significant skill for multi-year averages. Although not statistically significant, the decadal predictions tend to show a positive correlation with observations for the predictions of the Blocking regime during summer systematically at different forecast periods.

The comparison between the decadal predictions and the historical simulations indicates that the model initialization does not significantly improve the skill in predicting the weather regimes' variability. In fact, the skill decreases for some cases, especially for some multi-year averages of the NAO− regimes during winter. The highest potential benefit due to initialization is found for the Blocking regime during summer, with a systematic tendency toward positive ACC differences between DCPP and HIST across different forecast periods, but such differences are not statistically significant. The lack of added value due to initialization might be due to inconsistencies between the model and the initial conditions used to initialize the predictions (Bilbao et al., 2021). Better predictions of the variations of the Blocking regime's frequency could anticipate periods with more frequency of extreme events like cold air outbreaks, heat-waves, floods and droughts (Christensen et al., 2013), as well as episodes of high pollution over European regions (Garrido-Perez et al., 2017; Ordóñez et al., 2017).

While there is mounting evidence that climate models seem to underestimate some of the predictive signals related to atmospheric circulation, in particular in the Atlantic sector (Scaife & Smith, 2018; Smith et al., 2019), a few previous studies showed some predictability for selected weather types at multi-annual to decadal time scales. For example, Athanasiadis et al. (2020) showed decadal prediction skill for the High Latitude Blocking and the NAO index during the winter season using a large ensemble (40 members from the Community Earth System Model-Decadal Prediction Large Ensemble [CESM-DPLE; Yeager et al., 2018]). They showed that such a large ensemble allows the predictable component of the atmospheric variability to emerge from the chaotic noise. Smith et al. (2020) also showed predictability for the NAO index using a large multi-model ensemble from CMIP5 and CMIP6 models and applying post-processing techniques for overcoming the low signal-to-noise ratio of the raw model output. However, we use a different definition based on an objective weather regime classification designed to classify each day into a certain flow regime (or into the unclassified cluster).

In order to assess possible factors that limit the skill in predicting the weather regimes' frequencies, the observed and simulated teleconnections between the North Atlantic SST and the weather regime frequencies have been computed. The observed teleconnections show that the occurrence frequencies of the different weather regimes are significantly correlated with the SST over large regions of the North Atlantic Ocean (except for the winter Blocking regime, for which no significant correlations are found). However, these relationships are not shown in the teleconnections maps obtained with both the decadal predictions and historical simulations, limiting the skill that may potentially be transferred from the North Atlantic SST to the weather regimes' frequencies. Besides, the sea level pressure (variable used to compute the weather regimes) over the North Atlantic region particularly suffers from the signal-to-noise paradox (Smith et al., 2019), which may have contributed to reducing the skill in predicting the weather regimes. Previous studies have pointed to the benefits of using larger ensembles for predicting variables with a small predictive signal relative to the noise (Athanasiadis et al., 2020; Smith et al., 2020). In order to assess whether the results obtained with the relatively small ensemble size (10 ensemble members) can be improved using a larger ensemble, we have tested whether the skill improves when the ensemble size is doubled by including another 10 ensemble members with slightly different initialization but overall similar behavior (i.e., the skill is estimated with a total of 20 decadal prediction members), but the results are similar and no skill improvement is found (not shown). On the other hand, the low skill of the EC-Earth3 model in predicting the Subpolar North Atlantic SST shown by Bilbao et al. (2021) may also limit the skill that may potentially be transferred from the North Atlantic SST to the weather regimes' frequencies. The limitation of the skill in predicting the weather regimes' frequencies of occurrence due to biases in SSTs was also suggested by Fabiano et al. (2020). Another possible limitation might be the spatial resolution of the model. Recent works have shown that the representation of climatological frequencies and spatial patterns of some weather regimes (e.g., Blocking regime) can be improved by increasing the grid resolution of the atmospheric model (Dawson & Palmer, 2015; Fabiano et al., 2020). However, it is not clear if the increase in resolution can improve the prediction skill of the frequencies' variability.

In conclusion, this study demonstrates that the EC-Earth3 model, which is used for the contributions to CMIP6/DCPP-A, skillfully simulates most climatological aspects of Euro-Atlantic weather regimes. However, the skill in predicting the inter-annual to decadal variability of these weather regimes is low, and the model initialization does not significantly improve such skill (as seen by comparing the decadal predictions and historical simulations). This work can be the basis for more detailed future studies. For instance, further comparison with other models could help answer the question of whether the lack of skill in the prediction of the weather regimes frequencies by the EC-Earth3 model is due to an inherently unpredictable signal or due to model deficiencies (since other models might be skillful in predicting the temporal variations of the weather regimes). On the other hand, the multi-model ensemble of predictions contributing to CMIP6/DCPP-A could be used to assess if the skill is improved compared to the individual models due to error compensation and to the signal that each model adds to the multi-model ensemble (Hagedorn et al., 2005). Also, post-processing techniques such as the calibration method proposed by Eade et al. (2014) or the post-processing and member selection as introduced by Smith et al. (2020) for NAO predictions could be implemented with the goal to improve the prediction of the objectively classified weather regimes. If the predictability of weather regimes can eventually be established robustly in the future, it could unlock the potential for skillful decadal climate predictions over Europe and the prediction of specific weather phenomena, including extreme events typically related to certain weather regimes.

Data Availability Statement

The EC-Earth3 decadal predictions (members r[1–10]i1p1f1; EC-Earth, 2019b) and historical simulations (members r[2,12,14,16–18,21-22,24–25]i1p1f1; EC-Earth, 2019a) are available for downloading on the ESGF node (https://esgf-node.llnl.gov/search/cmip6/). JRA-55 reanalysis data (Japan Meteorological Agency, Japan, 2013) have been retrieved from https://climatedataguide.ucar.edu/climate-data/jra-55. NCEP reanalysis data (National Centers for Environmental Prediction et al., 1994) have been retrieved from https://www.esrl.noaa.gov/psd/. The code used in this paper can be found at https://earth.bsc.es/gitlab/cdelgado/cdelgado_copernicus/-/tree/development_branch/WeatherRegimes/paper_i1. We acknowledge the use of the s2dv, startR, multiApply, CSTools (Pérez-Zanón et al., 2021), and ClimProjDiags R software packages, all of them available on CRAN (https://cran.r-project.org/).

Acknowledgments

This research has received support by the AXA Research Fund, the CLINSA project (CGL2017-85791-R), and the EUCP project (Horizon 2020 Grant 776613). CDT acknowledges financial support from the Spanish Ministry for Science and Innovation (FPI PRE2019-08864 financed by MCIN/AEI/10.13039/501100011033 and by FSE invierte en tu futuro). EH was supported by the Spanish Project PRE4CAST (grant CGL2017-86415-R). MGD has also been supported by the Spanish Ministry for the Economy, Industry and Competitiveness (grant RYC-2017-22964). The authors thank Verónica Torralba, Nicola Cortesi, Roberto Bilbao, Lluís Palma, and Francisco J. Doblas-Reyes for their technical and scientific support. The authors also thank the anonymous reviewers for their valuable comments and suggestions, which improved the manuscript.

References

Athanasiadis, P. J., Yeager, S., Kwon, Y. O., Bellucci, A., Smith, D. W., & Tibaldi, S. (2020). Decadal predictability of North Atlantic blocking and the NAO. *npj Climate and Atmospheric Science*, *3*(1), 20. https://doi.org/10.1038/s41612-020-0120-6

Bilbao, R., Wild, S., Ortega, P., Acosta-Navarro, J., Arsouze, T., Bretonnière, P. A., et al. (2021). Assessment of a full-field initialized decadal climate prediction system with the CMIP6 version of EC-Earth. *Earth System Dynamics*, *12*(1), 173–196. https://doi.org/10.5194/esd-12-173-2021

Boer, G. J., Smith, D. M., Cassou, C., Doblas-Reyes, F., Danabasoglu, G., Kirtman, B., et al. (2016). The Decadal Climate Prediction Project (DCPP) contribution to CMIP6. *Geoscientific Model Development*, *9*(10), 3751–3777. https://doi.org/10.5194/gmd-9-3751-2016

Carvalho-Oliveira, J., Borchert, L. F., Zorita, E., & Baehr, J. (2022). Self-organizing maps identify windows of opportunity for seasonal European summer predictions. *Frontiers in Climate*, *4*, 30. https://doi.org/10.3389/FCLIM.2022.844634/BIBTEX

Casola, J. H., & Wallace, J. M. (2007). Identifying weather regimes in the wintertime 500-hPa geopotential height field for the Pacific–North American sector using a limited-contour clustering technique. *Journal of Applied Meteorology and Climatology*, *46*(10), 1619–1630. https://doi.org/10.1175/JAM2564.1

Cassou, C., Terray, L., & Phillips, A. S. (2005). Tropical Atlantic influence on European heat waves. *Journal of Climate*, *18*(15), 2805–2811. https://doi.org/10.1175/JCLI3506.1

Cattiaux, J., Douville, H., & Peings, Y. (2013). European temperatures in CMIP5: Origins of present-day biases and future uncertainties. *Climate Dynamics*, *41*(11–12), 2889–2907. https://doi.org/10.1007/s00382-013-1731-y

Cattiaux, J., Quesada, B., Arakélian, A., Codron, F., Vautard, R., & Yiou, P. (2013). North-Atlantic dynamics and European temperature extremes in the IPSL model: Sensitivity to atmospheric resolution. *Climate Dynamics*, *40*(9–10), 2293–2310. https://doi.org/10.1007/s00382-012-1529-3

Christensen, J. H., Kanikicharla, K. K., Aldrian, E., An, S. I., Albuquerque Cavalcanti, I. F., de Castro, M., et al. (2013). *Climate phenomena and their relevance for future regional climate change*. In *Climate change 2013 the physical science basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. https://doi.org/10.1017/CBO9781107415324.028

Christiansen, B. (2007). Atmospheric circulation regimes: Can cluster analysis provide the number? *Journal of Climate*, *20*(10), 2229–2250. https://doi.org/10.1175/JCLI4107.1

Cleveland, W. S., & Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, *83*(403), 596–610. https://doi.org/10.1080/01621459.1988.10478639

Cortesi, N., González-Reviriego, N., Soret, A., & Doblas-Reyes, F. J. (2017). *Weather regimes: ECMWF seasonal forecasts verification*. (Technical Report). Barcelona Supercomputing Center - Centro Nacional de Supercomputación (BSC-CNS). Retrieved from https://earth.bsc.es/wiki/lib/exe/fetch.php?media=library:external:20170404%7B%5C_%7Dncortesi%7B%5C_%7Dseasonal%7B%5C_%7Dforecasts%7B%5C_%7D-verification.pdf

Cortesi, N., Torralba, V., González-Reviriego, N., Soret, A., & Doblas-Reyes, F. J. (2019). Characterization of European wind speed variability using weather regimes. *Climate Dynamics*, *53*(7–8), 4961–4976. https://doi.org/10.1007/s00382-019-04839-5

Cortesi, N., Torralba, V., Lledó, L., Manrique-Suñén, A., Gonzalez-Reviriego, N., Soret, A., & Doblas-Reyes, F. J. (2021). Yearly evolution of Euro-Atlantic weather regimes and of their sub-seasonal predictability. *Climate Dynamics*, *56*(11–12), 3933–3964. https://doi.org/10.1007/s00382-021-05679-y

D'Andrea, F., Tibaldi, S., Blackburn, M., Boer, G., Déqué, M., Dix, M. R., et al. (1998). Northern Hemisphere atmospheric blocking as simulated by 15 atmospheric general circulation models in the period 1979–1988. *Climate Dynamics*, *14*(6), 385–407. https://doi.org/10.1007/s003820050230

Dawson, A., & Palmer, T. N. (2015). Simulating weather regimes: Impact of model resolution and stochastic parameterization. *Climate Dynamics*, *44*(7), 2177–2193. https://doi.org/10.1007/S00382-014-2238-X

Dawson, A., Palmer, T. N., & Corti, S. (2012). Simulating regime structures in weather and climate prediction models. *Geophysical Research Letters*, *39*(21), 21805. https://doi.org/10.1029/2012GL053284

Doblas-Reyes, F. J., Andreu-Burillo, I., Chikamoto, Y., García-Serrano, J., Guemas, V., Kimoto, M., et al. (2013). Initialized near-term regional climate change prediction. *Nature Communications*, *4*(1), 1715. https://doi.org/10.1038/ncomms2704

Donat, M. G., Leckebusch, G. C., Pinto, J. G., & Ulbrich, U. (2010). European storminess and associated circulation weather types: Future changes deduced from a multi-model ensemble of GCM simulations. *Climate Research*, *42*(1), 27–43. https://doi.org/10.3354/cr00853

Döscher, R., Acosta, M., Alessandri, A., Anthoni, P., Arsouze, T., Bergman, T., et al. (2022). The EC-Earth3 Earth system model for the Coupled Model Intercomparison Project 6. *Geoscientific Model Development*, *15*(7), 2973–3020. https://doi.org/10.5194/GMD-15-2973-2022

Eade, R., Smith, D., Scaife, A., Wallace, E., Dunstone, N., Hermanson, L., & Robinson, N. (2014). Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophysical Research Letters*, *41*(15), 5620–5628. https://doi.org/10.1002/2014GL061146

EC-Earth. (2019a). EC-Earth-Consortium EC-Earth3 model output prepared for CMIP6 CMIP historical [dataset]. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.4700

EC-Earth. (2019b). EC-Earth-Consortium EC-Earth3 model output prepared for CMIP6 DCPP DCPPA-hindcast [dataset]. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.4553

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Fabiano, F., Christensen, H. M., Strommen, K., Athanasiadis, P., Baker, A., Schiemann, R., & Corti, S. (2020). Euro-Atlantic weather Regimes in the PRIMAVERA coupled climate simulations: Impact of resolution and mean state biases on model performance. *Climate Dynamics*, *54*(11), 5031–5048. https://doi.org/10.1007/S00382-020-05271-W

Falkena, S. K., de Wiljes, J., Weisheimer, A., & Shepherd, T. G. (2020). Revisiting the identification of wintertime atmospheric circulation regimes in the Euro-Atlantic sector. *Quarterly Journal of the Royal Meteorological Society*, *146*(731), 2801–2814. https://doi.org/10.1002/qj.3818

Fereday, D. R., Knight, J. R., Scaife, A. A., Folland, C. K., & Philipp, A. (2008). Cluster analysis of North Atlantic-European circulation types and links with tropical Pacific sea surface temperatures. *Journal of Climate*, *21*(15), 3687–3703. https://doi.org/10.1175/2007JCLI1875.1

Ferranti, L., Corti, S., & Janousek, M. (2015). Flow-dependent verification of the ECMWF ensemble over the Euro-Atlantic sector. *Quarterly Journal of the Royal Meteorological Society*, *141*(688), 916–924. https://doi.org/10.1002/qj.2411

Fil, C., & Dubus, L. (2005). Winter climate regimes over the North Atlantic and European region in ERA40 reanalysis and DEMETER seasonal hindcasts. *Tellus, Series A: Dynamic Meteorology and Oceanography*, *57*(3), 290–307. https://doi.org/10.1111/j.1600-0870.2005.00127.x

Garrido-Perez, J. M., Ordóñez, C., & García-Herrera, R. (2017). Strong signatures of high-latitude blocks and subtropical ridges in winter PM_{10} over Europe. *Atmospheric Environment*, *167*, 49–60. https://doi.org/10.1016/j.atmosenv.2017.08.004

Giorgetta, M. A., Jungclaus, J., Reick, C. H., Legutke, S., Bader, J., Böttinger, M., et al. (2013). Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project phase 5. *Journal of Advances in Modeling Earth Systems*, *5*(3), 572–597. https://doi.org/10.1002/JAME.20038

Hagedorn, R., Doblas-Reyes, F. J., & Palmer, T. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept. *Tellus A: Dynamic Meteorology and Oceanography*, *57*(3), 219–233. https://doi.org/10.3402/TELLUSA.V57I3.14657

Hertig, E., & Jacobeit, J. (2014). Variability of weather regimes in the North Atlantic-European area: Past and future. *Atmospheric Science Letters*, *15*(4), 314–320. https://doi.org/10.1002/asl2.505

Japan Meteorological Agency, Japan. (2013). JRA-55: Japanese 55-year reanalysis, daily 3-hourly and 6-hourly data [dataset]. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. https://doi.org/10.5065/D6HH6H41

Kageyama, M., D'Andrea, F., Ramstein, G., Valdes, P. J., & Vautard, R. (1999). Weather regimes in past climate atmospheric general circulation model simulations. *Climate Dynamics*, *15*(10), 773–793. https://doi.org/10.1007/S003820050315

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., et al. (1996). The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, *77*(3), 437–471. https://doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2

Kirtman, B., Power, S. B., Adedoyin, A. J., Boer, G. J., Bojariu, R., Camilloni, I., et al. (2013). *Near-term climate change: Projections and predictability*. In *Climate change 2013 the physical science basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. https://doi.org/10.1017/CBO9781107415324.023

Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., et al. (2015). The JRA-55 reanalysis: General specifications and basic characteristics. *Journal of the Meteorological Society of Japan*, *93*(1), 5–48. https://doi.org/10.2151/jmsj.2015-001

Kushnir, Y., Scaife, A. A., Arritt, R., Balsamo, G., Boer, G., Doblas-Reyes, F., et al. (2019). Towards operational predictions of the near-term climate. *Nature Climate Change*, *9*(2), 94–101. https://doi.org/10.1038/s41558-018-0359-7

Mariotti, A., Baggett, C., Barnes, E. A., Becker, E., Butler, A., Collins, D. C., et al. (2020). Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond. *Bulletin of the American Meteorological Society*, *101*(5), E608–E625. https://doi.org/10.1175/BAMS-D-18-0326.1

Masato, G., Hoskins, B. J., & Woollings, T. (2013). Winter and summer Northern Hemisphere blocking in CMIP5 models. *Journal of Climate*, *26*(18), 7044–7059. https://doi.org/10.1175/JCLI-D-12-00466.1

Meehl, G. A., Goddard, L., Murphy, J., Stouffer, R. J., Boer, G., Danabasoglu, G., et al. (2009). Decadal prediction. *Bulletin of the American Meteorological Society*, *90*(10), 1467–1486. https://doi.org/10.1175/2009bams2778.1

Meehl, G. A., Richter, J. H., Teng, H., Capotondi, A., Cobb, K., Doblas-Reyes, F., et al. (2021). Initialized Earth system prediction from subseasonal to decadal timescales. *Nature Reviews Earth & Environment*, *2*(5), 340–357. https://doi.org/10.1038/s43017-021-00155-x

Merryfield, W. J., Baehr, J., Batté, L., Becker, E. J., Butler, A. H., Coelho, C. A., et al. (2020). Current and emerging developments in subseasonal to decadal prediction. *Bulletin of the American Meteorological Society*, *101*(6), E869–E896. https://doi.org/10.1175/BAMS-D-19-0037.1

Michelangeli, P.-A., Vautard, R., & Legras, B. (1995). Weather regimes: Recurrence and quasi stationarity. *Journal of the Atmospheric Sciences*, *52*(8), 1237–1256. https://doi.org/10.1175/1520-0469(1995)052<1237:WRRAQS>2.0.CO;2

Molteni, F., Stockdale, T., Balmaseda, M., Balsamo, G., Buizza, R., Ferranti, L., et al. (2011). The new ECMWF seasonal forecast system (System 4). *ECMWF technical report*.

National Centers for Environmental Prediction, National Weather Service, NOAA, & U.S. Department of Commerce. (1994). NCEP/NCAR global reanalysis products, 1948-continuing [dataset]. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. Retrieved from https://rda.ucar.edu/datasets/ds090.0/

Ordóñez, C., Barriopedro, D., Garciá-Herrera, R., Sousa, P. M., & Schnell, J. L. (2017). Regional responses of surface ozone in Europe to the location of high-latitude blocks and subtropical ridges. *Atmospheric Chemistry and Physics*, *17*(4), 3111–3131. https://doi.org/10.5194/acp-17-3111-2017

Pérez-Zanón, N., Caron, L.-P., Terzago, S., Schaeybroeck, B. V., Lledó, L., Manubens, N., et al. (2021). The CSTools (v4.0) toolbox: From climate forecasts to climate forecast information. *Geoscientific Model Development Discussions*, *2021*, 1–32. https://doi.org/10.5194/gmd-2021-368

Philipp, A., Bartholy, J., Beck, C., Erpicum, M., Esteban, P., Fettweis, X., et al. (2010). Cost733cat – A database of weather and circulation type classifications. *Physics and Chemistry of the Earth*, *35*(9–12), 360–373. https://doi.org/10.1016/j.pce.2009.12.010

Rodwell, M. J., Rowell, D. P., & Folland, C. K. (1999). Oceanic forcing of the wintertime North Atlantic Oscillation and European climate. *Nature*, *398*(6725), 320–323. https://doi.org/10.1038/18648

Scaife, A. A., & Smith, D. (2018). A signal-to-noise paradox in climate science. *npj Climate and Atmospheric Science*, *1*, 28. https://doi.org/10.1038/s41612-018-0038-4

Schiemann, R., Demory, M. E., Shaffrey, L. C., Strachana, J., Vidale, P. L., Mizielinski, M. S., et al. (2017). The resolution sensitivity of Northern Hemisphere blocking in four 25-km atmospheric global circulation models. *Journal of Climate*, *30*(1), 337–358. https://doi.org/10.1175/JCLI-D-16-0100.1

Smith, D. M., Eade, R., Scaife, A. A., Caron, L. P., Danabasoglu, G., DelSole, T. M., et al. (2019). Robust skill of decadal climate predictions. *npj Climate and Atmospheric Science*, *2*(1), 13. https://doi.org/10.1038/s41612-019-0071-y

Smith, D. M., Scaife, A. A., Eade, R., Athanasiadis, P., Bellucci, A., Bethke, I., et al. (2020). North Atlantic climate far more predictable than models imply. *Nature*, *583*(7818), 796–800. https://doi.org/10.1038/s41586-020-2525-0

Stephenson, D. B., Hannachi, A., & O'Neill, A. (2004). On the existence of multiple climate regimes. *Quarterly Journal of the Royal Meteorological Society*, *130*(597), 583–605. https://doi.org/10.1256/QJ.02.146

Stryhal, J., & Huth, R. (2017). Classifications of winter Euro-Atlantic circulation patterns: An intercomparison of five atmospheric reanalyses. *Journal of Climate*, *30*(19), 7847–7861. https://doi.org/10.1175/JCLI-D-17-0059.1

Sutton, R. T., & Allen, M. R. (1997). Decadal predictability of North Atlantic sea surface temperature and climate. *Nature*, *388*(6642), 563–567. https://doi.org/10.1038/41523

Torralba, V. (2019). *Seasonal climate prediction for the wind energy sector: Methods and tools for the development of a climate service* (Doctoral dissertation). Complutense University of Madrid.

Torralba, V., Gonzalez-Reviriego, N., Cortesi, N., Manrique-Suñén, A., Lledó, L., Marcos, R., et al. (2021). Challenges in the selection of atmospheric circulation patterns for the wind energy sector. *International Journal of Climatology*, *41*(3), 1525–1541. https://doi.org/10.1002/joc.6881

Walz, M. A., Donat, M. G., & Leckebusch, G. C. (2018). Large-scale drivers and seasonal predictability of extreme wind speeds over the North Atlantic and Europe. *Journal of Geophysical Research: Atmospheres*, *123*(20), 11518–11535. https://doi.org/10.1029/2017JD027958

Wilks, D. S. (2011). Forecast verification. *International Geophysics*, *100*, 301–394. https://doi.org/10.1016/B978-0-12-385022-5.00008-7

Yeager, S. G., Danabasoglu, G., Rosenbloom, N. A., Strand, W., Bates, S. C., Meehl, G. A., et al. (2018). Predicting near-term changes in the Earth system: A large ensemble of initialized decadal prediction simulations using the Community Earth System Model. *Bulletin of the American Meteorological Society*, *99*(9), 1867–1886. https://doi.org/10.1175/BAMS-D-17-0098.1

Zwiers, F. W., & von Storch, H. (1995). Taking serial correlation into account in tests of the mean. *Journal of Climate*, *8*(2), 336–351. https://doi.org/10.1175/1520-0442(1995)008<0336:TSCIAI>2.0.CO;2

Chapter 5

Multi-annual predictions of the frequency and intensity of daily temperature and precipitation extremes

This chapter has been published as peer-reviewed article as:

Delgado-Torres, C., Donat, M. G., Soret, A., Gonzalez-Reviriego, N., Bretonnière, P.-A., Ho, A.-C., Pérez-Zanón, N., Samsó Cabré, M., and Doblas-Reyes, F. J. (2023). Multi-annual predictions of the frequency and intensity of daily temperature and precipitation extremes. Environmental Research Letters, 18 034031. https://doi.org/10.1088/1748-9326/acbbe1

The supplementary material can be found in Appendix D.

5.1. Main objectives

- Estimate the multi-model forecast quality for extreme indices based on daily minimum and maximum temperature and precipitation at multi-annual time scales.

- Compare the skill in predicting temperature and precipitation extremes to that for mean temperature and precipitation.
- Estimate the impact of model initialisation on the skill in predicting climate extremes.

5.2. Main outcomes

- The multi-model ensemble skillfully predicts the temperature extremes over most land regions, while the prediction skill is more limited for precipitation extremes.
- The extreme indices are predicted with lower skill than the mean quantities.
- The skill in predicting extreme indices based on minimum temperature is generally higher than that for indices based on maximum temperature.
- The extreme indices based on percentiles are predicted with higher skill than those representing the most extreme days.
- The added value from model initialisation is generally low and highly region-dependent.

ENVIRONMENTAL RESEARCH
LETTERS

OPEN ACCESS

RECEIVED
27 October 2022

REVISED
24 January 2023

ACCEPTED FOR PUBLICATION
14 February 2023

PUBLISHED
28 February 2023

LETTER

Multi-annual predictions of the frequency and intensity of daily temperature and precipitation extremes

Carlos Delgado-Torres[1,*], Markus G Donat[1,2], Albert Soret[1], Nube González-Reviriego[1], Pierre-Antoine Bretonnière[1], An-Chi Ho[1], Núria Pérez-Zanón[1], Margarida Samsó Cabré[1] and Francisco J Doblas-Reyes[1,2]

[1] Barcelona Supercomputing Center (BSC), Barcelona, Spain
[2] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain
[*] Author to whom any correspondence should be addressed.

E-mail: carlos.delgado@bsc.es

Keywords: decadal climate predictions, extreme climate events, forecast quality assessment, CMIP6 multi-model ensemble, impact of model initialization

Supplementary material for this article is available online

Abstract

The occurrence of extreme climate events in the coming years is modulated by both global warming and internal climate variability. Anticipating changes in frequency and intensity of such events in advance may help minimize the impact on climate-vulnerable sectors and society. Decadal climate predictions have been developed as a source of climate information relevant for decision-making at multi-annual timescales. We evaluate the multi-model forecast quality of the CMIP6 decadal hindcasts in predicting a set of indices measuring different characteristics of temperature and precipitation extremes for the forecast years 1–5. The multi-model ensemble skillfully predicts the temperature extremes over most land regions, while the skill is more limited for precipitation extremes. We further compare the prediction skill for these extreme indices to the skill for mean temperature and precipitation, finding that the extreme indices are predicted with lower skill, particularly those representing the most extreme days. We find only small and region-dependent improvements from model initialization in comparison to historical forcing simulations. This systematic evaluation of decadal hindcasts is essential when providing a climate service based on decadal predictions so that the user is informed on the trustworthiness of the forecasts for each specific region and extreme event.

1. Introduction

Characteristics of climate extremes are changing in a warming climate, with in particular hot temperature and heavy precipitation extremes becoming more intense and frequent, thus increasing their potential impact on nature, economy and society (Seneviratne *et al* 2021). Besides, internal climate variability also modulates the occurrence of extreme events (Alexander *et al* 2009). Trustworthy predictions are essential to develop strategic planning to adapt, build more resilience to the risk associated with extreme events and anticipate the impacts ahead of time (Hanlon *et al* 2013, Curtis *et al* 2017, Sillmann *et al* 2017, Kushnir *et al* 2019). Predictions of changes in the frequency and intensity of extreme events may be more relevant to users than predictions of average variables, as the occurrence of extremes typically breaks the resilience of a system and cause the heaviest impacts on society and environment (Mahlstein *et al* 2015, Bhend *et al* 2017).

Predictions of variations in the frequency and intensity of extremes in the forthcoming years can potentially be provided by decadal climate predictions. In addition to long-term changes due to external forcings (natural and anthropogenic), decadal predictions aim to also capture the internal variability of the climate system (slow, natural oscillations). For this reason, climate models are initialized with observation-based products (Meehl *et al* 2009, 2021). Decadal predictions have been shown to skillfully predict essential climate variables such as near-surface air temperature and, to a lesser extent, precipitation (Smith *et al* 2019, Delgado-Torres *et al* 2022) in many regions of the world. However, the predictability of mean quantities and extreme events might differ (Liu *et al* 2019).

Only a few studies have previously evaluated temperature and precipitation extremes in decadal predictions. Eade *et al* (2012) found a generally high skill for temperature extremes, and limited skill over parts of North America for precipitation extremes with the Met Office Decadal Prediction System (DePreSys) (Smith *et al* 2007, 2010). Also, they found a slight skill improvement due to initialization beyond the first year, pointing to external forcings as the primary source of skill. Besides, they found slightly lower skill for extreme predictions than mean quantities, except for those regions with a greater trend in extremes than in the mean variables. Hanlon *et al* (2013) also found significant skill for multi-year predictions of summer temperature extremes with DePreSys over Europe, while the skill for winter extremes was lower. They also found lower skill in predicting temperature extremes than mean quantities. Decadal predictions from four models contributed to the Coupled Model Intercomparison Project Phase 5 (CMIP5) (Taylor *et al* 2012) were evaluated by Hanlon *et al* (2013) for European summer extremes, showing higher skill than the predictions based on observed climatology. Also, they found that the skill did not improve with initialization, except for one of the models. DePreSys has also been shown to outperform climatology and persistence forecasts for a set of daily temperature extreme indices over parts of Europe during summer for 10-year predictions (Hanlon *et al* 2015). Nevertheless, the model initialization did not improve the forecast quality. High skill for European temperature extremes and lower skill for precipitation extremes was also found with the Mittelfristige Klimaprognosen (German term for 'mid-term climate forecast'; MiKlip) system (Moemken *et al* 2021).

However, to our best knowledge, all previous studies were based on single forecast systems or limited regions to evaluate the decadal forecast quality for climate extremes. We perform a forecast quality assessment of multi-model decadal predictions of annual and seasonal extreme temperature and precipitation indices with all available hindcasts contributed to the Decadal Climate Prediction Project Component A (DCPP-A) (Boer *et al* 2016) of the Coupled Model Intercomparison Project Phase 6 (CMIP6) (Eyring *et al* 2016). The evaluation is performed globally for predictions of the next five years. The skill for extreme predictions is compared to that for mean temperature and precipitation variations, and the impact of model initialization is assessed by comparing the skill of decadal predictions and historical forcing simulations.

2. Data

Daily minimum and maximum near-surface air temperature and precipitation have been used to compute the extreme indices. Besides, monthly means of near-surface air temperature (TAS) and precipitation (PR) have been used to compare the prediction skill for extreme indices and mean variables.

All the available CMIP6 decadal hindcasts (DCPP) have been used. Besides, the CMIP6 historical simulations (HIST) performed with the same forecast systems as DCPP have been used to estimate the impact of the model initialization on the prediction skill. The number of DCPP and HIST members provided by each forecast system and their basic information can be seen in table S1. The DCPP and HIST multi-model ensembles consist of 133 and 134 members, respectively.

Two different gridded observation-based datasets per variable have been used as the reference for the evaluation (table S2) to account for the observational uncertainty (Sillmann *et al* 2013, Alexander *et al* 2020). The Berkeley Earth Surface Temperatures (BEST) and Rainfall Estimates on a Gridded Network (REGEN) (Contractor *et al* 2020) datasets provide gridded fields of daily minimum and maximum temperatures and daily PR, respectively. These have been used to calculate the extremes indices and are used as observational references. The HadEX3 (Dunn *et al* 2020) dataset (which provides gridded extremes indices calculated for each station and then interpolated onto global grids) has been used as an additional reference dataset. The results obtained with BEST and REGEN are shown in the main text, while those obtained with HadEX3 are shown in the Supplementary Material. The Global Historical Climatology Network version 4 (GHCNv4) (Menne *et al* 2018) and Global Precipitation Climatology Centre (GPCC) (Schneider *et al* 2020) datasets have been used as references for TAS and PR, respectively.

3. Methods

The Expert Team on Climate Change Detection and Indices (ETCCDI) defined a set of extreme climate indices to detect, characterize and monitor changes in the frequency and severity of extreme events, such as heat waves, cold spells, floods and droughts (Zhang *et al* 2011). We have selected six extreme indices:

- TN10p: seasonal or annual percentage of days when minimum temperature is below the 10th daily percentile.
- TNn: seasonal or annual minimum of daily minimum temperature.
- TX90p: seasonal or annual percentage of days when maximum temperature is above the 90th daily percentile.

- TXx: seasonal or annual maximum of daily maximum temperature.
- R95p: annual sum of precipitation in days where daily PR exceeds the 95th percentile of daily precipitation.
- Rx5day: seasonal or annual maximum 5-day consecutive precipitation.

Thus, for each variable, we evaluate a measure of relatively moderate extremes, which occur on average several times per year or season, and a measure representing the most intense event of the year or season. The R-based software package we use to compute the ETCCDI indices (climdex.pcic) (Bronaugh 2020) provides the TN10p, TNn, TX90p, TXx, and Rx5day indices at seasonal and annual frequencies, while it only provides the R95p index at annual aggregation. Therefore, all six indices have been evaluated at annual frequency. Besides, the boreal winter (DJF; December-January-February) and boreal summer (JJA; June-July-August) indices that can also be computed at seasonal frequency have been included in the analysis. Similarly, annual and seasonal averages of TAS and PR have been analyzed.

The extreme indices have been evaluated globally during the 1961–2014 period using the 1981–2010 period as the baseline period for the percentile-based indices calculation. The same reference period has been used to compute the climatology and thresholds between the tercile probabilistic categories. For DCPP, the average of the forecast years 1 to 5 has been evaluated. Thus, start dates 1960–2009 have been used. In the case of the forecast systems not initialized in January (see table S1), the first forecast months have been discarded to define the analysis to calendar years (i.e. from January to December). A 5-year running mean has been applied to HIST and reference datasets to make a consistent comparison with the forecast period of DCPP.

The multi-model ensembles have been built by pooling all members together. The extreme indices have been computed in the native grid of each dataset (see tables S1 and S2) to avoid smoothing the extreme values when interpolating daily fields. For the evaluation, the mean variables and extreme indices have been interpolated to a common 2.8° × 2.8° grid resolution using the conservative interpolation method.

Both deterministic and probabilistic predictions have been evaluated. The deterministic forecasts are based on the multi-model ensemble mean, while the probabilistic forecasts are based on the percentage of ensemble members that fall into each tercile category (below lower tercile, near average, and above upper tercile conditions). The tercile categories have been estimated based on the 33.33% and 66.67% thresholds of the corresponding probability density functions.

The Spearman's anomaly correlation coefficient (ACC) (Wilks 2011), which estimates the linear relationship between the observed and predicted time series, has been used to evaluate the deterministic forecasts. Spearman's correlation has been chosen to avoid assuming that the data are normally distributed. We have also tested the sensitivity of using Pearson's correlation coefficient instead of Spearman's, finding correlation values generally very similar. The ACC ranges between −1 (worst forecast) and 1 (perfect forecast). The residual correlation (Smith *et al* 2019) has been used to assess whether DCPP predicts any of the observed variability that is not already captured by HIST forced signal, and also ranges from −1 to 1. The residual correlation is computed as follows: the residuals of both the DCPP ensemble mean and reference dataset are calculated by linearly regressing out the HIST ensemble mean from the DCPP ensemble mean and reference dataset, respectively. The residual correlation is computed as the correlation between both residuals. Positive (negative) values indicate that DCPP predicts the observed variability better (worse) than HIST.

The quality of the probabilistic predictions has been evaluated with the ranked probability skill score (RPSS) (Wilks 2011), which measures the quality of a forecast in comparison with a reference forecast, and ranges between minus infinity and 1. Negative values indicate that the reference forecast is more skillful than the forecast, while positive values mean the opposite. The DCPP forecasts have been compared to the climatological forecast (defined as the equiprobable forecast, with a probability of 33.33% for each tercile category) and to the HIST.

The statistical significance of the ACC has been estimated with a one-sided t-test (Wilks 2011) accounting for the time series auto-correlation following Zwiers and von Storch (1995) with the null hypothesis that the ACC is not positive. We use a one-sided test as only positive correlation values carry useful predictive information. The same test but two-sided (in order to identify both potential improvements and deteriorations from initialization) has been applied for the statistical significance of the residual correlation with the null hypothesis that the residual correlation equals zero. The statistical significance of the RPSS using the climatological forecast as the reference forecast has been estimated with a one-sided Random Walk test (Delsole and Tippett 2016) with the null hypothesis that DCPP has less than or equal to 50% probability of being more skillful than the climatological forecast. The same test but two-sided has been applied for the statistical significance of the RPSS using HIST as the reference forecast with the null hypothesis that DCPP has a probability different than 50% of being more or less skillful than HIST. We control for multiple testing by applying

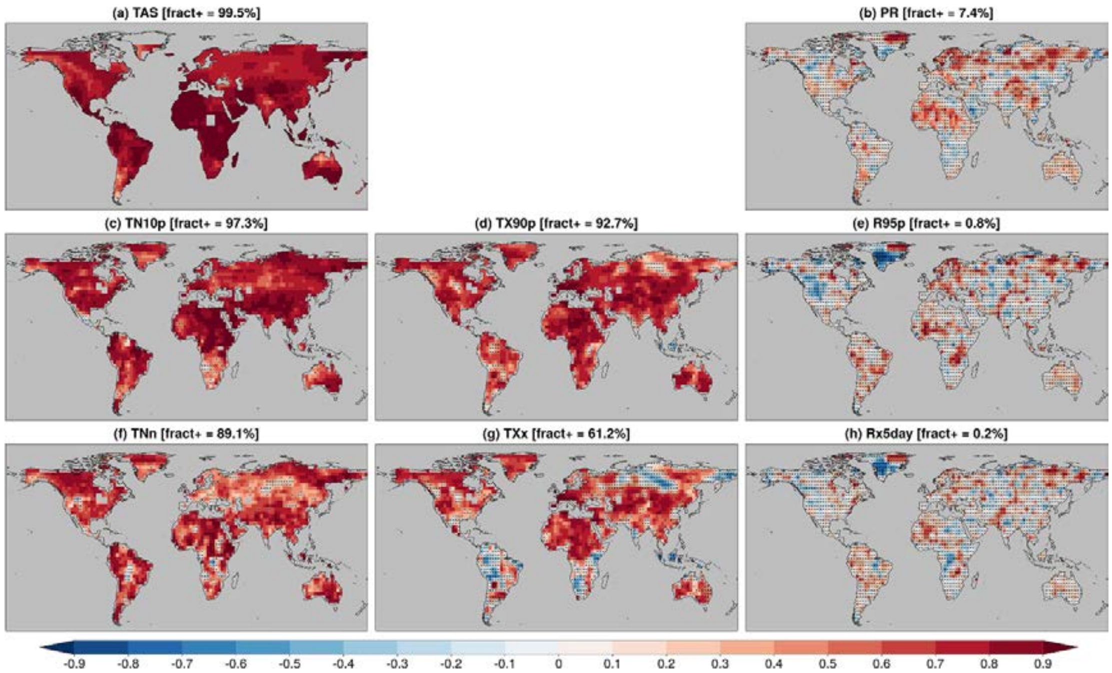


Figure 1. Maps of ACC obtained with the DCPP multi-model ensemble for the forecast years 1–5 (annual means) for the mean variables and extreme indices. The percentage of the global area with statistically significant positive or negative values is shown in the titles. The reference datasets used for the mean near-surface air temperature and PR are the GHCNv4 and the GPCC datasets, respectively. The reference datasets used for the temperature and PR indices are the BEST and REGEN datasets, respectively. Grey colors over land regions correspond to those grid points with missing values in the reference dataset. Crosses indicate that the values are not statistically significant using a one-sided t-test accounting for autocorrelation and controlling the FDR with $\alpha_{FDR} = 0.1$.

the (Wilks 2016) false detection rate (FDR) procedure using $\alpha_{FDR} = 0.1$, which approximately corresponds to a global significance level of $\alpha = 0.05$. The deterministic evaluation is shown in the main text, and the probabilistic evaluation in the supplementary material.

4. Results and discussion

4.1. Multi-model skill for annual extreme indices

The DCPP shows high and significant skill in predicting TAS over most of the globe (figure 1(a)), with 99.5% of the global regions being statistically significant. The high and significant skill of the DCPP multi-model is consistent with the results for mean temperature and PR reported by Smith *et al* (2019) and Delgado-Torres *et al* (2022). For extreme temperature, the skill is generally lower, and significant skill is found over smaller areas of the globe than for TAS, especially for the indices representing the annual most extreme day (TNn and TXx; figures 1(f) and (g), respectively). The lower skill found for predictions of extremes than for TAS is consistent with Eade *et al* (2012). However, they found higher skill for extremes than for TAS in some cases. Specifically, they showed that DePreSys presents a higher skill for extremes in regions where trends in extremes are larger than for TAS (e.g. rainfall over Europe, hot extremes over northern Eurasia, and cold extremes over the USA). With the CMIP6 multi-model, we find opposite results for hot extremes. For example, predictions of hot extremes over northern Eurasia (e.g. TX90p; figure 1(d)) show less grid points with significant skill than for TAS (figure 1(a)). For cold extremes (TN10p; figure 1(c)), results are consistent with those reported by Eade *et al* (2012) over the USA, with some areas showing a higher skill for TN10p than for TAS. Liu *et al* (2019) found a potentially higher predictability for moderate temperature extremes than for TAS in a perfect-model experiment. However, the CMIP6 multi-model ensemble, initialized with observation-based initial conditions, shows a generally higher skill for TAS than for moderate extremes.

Still, the skill for hot and cold extremes is high and significant over many regions, particularly over some areas of North America, North Africa, and parts of Eurasia and Australia. The comparison between hot and cold extremes shows a generally higher skill and larger areas with significant skill for minimum temperature extremes (TN10p and TNn; figures 1(c) and (d), respectively) than those based on maximum temperature (TN90p and TXx; figures 1(d) and (g), respectively).

The skill for PR is more limited than for TAS, with 7.4% of the global region being significant (figure 1(b)). The regions in which DCPP are skillful in predicting PR are parts of Northern Africa, Northern Europe, Australia and Eurasia. As for temperature, the predictions of PR extremes are generally less skillful than for PR (figures 1(e) and (h)). Besides, the skill maps for PR extremes are noisier and positive significance is found over very limited regions.

The higher skill for extreme temperature predictions compared to extreme PR is consistent with previous studies evaluating predictions for mean variables (Smith *et al* 2019, Delgado-Torres *et al* 2022), and can be caused by several reasons. The primary source of prediction skill for TAS is the response to forcings, which causes a trend that climate models capture relatively well (Smith *et al* 2019). On the contrary, PR variability depends more on atmospheric circulation, which is often less well simulated by climate models (vanUlden and vanOldenborgh 2006). In addition, PR is more strongly affected by the signal-to-noise issue in climate models (Scaife and Smith 2018, Smith *et al* 2019), which makes PR a variable more challenging to predict. Furthermore, the low resolution of climate models does not allow small-scale processes, such as convection, to be resolved (Merryfield *et al* 2020). Therefore, a parametrization is needed, which also contributes as an error source. However, PR predictions are still skillful over some regions such as Europe and Sahelian Africa, and the Atlantic multidecadal variability may be the source of predictability for PR over these regions (Doblas-Reyes *et al* 2013).

The evaluation of probabilistic forecasts (figure S1) shows a benefit from using decadal predictions instead of climatology since positive RPSS values are found over most regions for TAS and temperature extremes. Low RPSS values are found over most of the globe for PR and PR extremes, indicating small or no benefit from using DCPP instead of the climatological forecast.

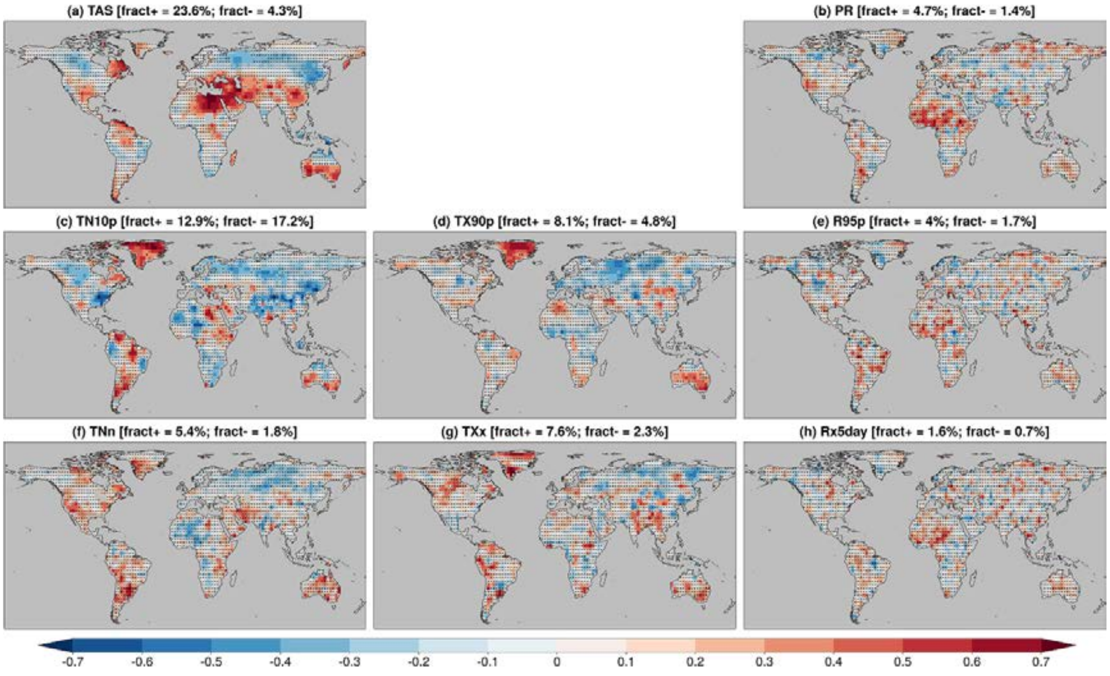
The results obtained with HadEX3 (figures S2 and S3) are generally consistent with the previously reported results, with similar skill values and significance for most indices and regions. The most noticeable exception is TN10p over South America and TX90p over Asia, for which the skill is higher and more positive significance is found when using HadEX3 instead of BEST as reference dataset. It should be noted that not all the regions can be compared due to the limited global coverage of HadEX3.

4.2. Impact of initialization for annual extreme indices

The residual correlations obtained for TAS show significant added skill from initialization over regions of Central America, North Africa, Eurasia and Southern Australia (figure 2(a)). Instead, DCPP show a significantly reduced skill for TAS predictions over the parts of Eurasia and Canada. Still, the global area with residual correlations showing positive significance is higher than negative significance (23.6% and 4.3%, respectively). For PR, the significant added skill is lower and mainly restricted over Central Africa (figure 2(b)). Nevertheless, there are more regions showing positive than negative significance (4.7% and 1.4%, respectively). These results are in line with what was reported by Smith *et al* (2019) and Delgado-Torres *et al* (2022) for TAS and PR. For temperature and PR extremes, the patterns of the impact of initialization that we find differs from those found with DePreSys (Eade *et al* 2012). The differences may be caused by using a multi-model ensemble instead of a single model, the different generations of prediction systems, the metric used to assess the impact of initialization, and the different forcings (which have been shown to provide more skill in CMIP6 than in CMIP5 (Borchert *et al* 2021)).

The initialization aims at phasing the simulations with the observed climate state. However, model initialization is a nontrivial procedure with various issues (Merryfield *et al* 2020) that can reduce the skill compared to HIST. For example, initialization shocks due to inconsistencies between the initial conditions and model climatology can degrade the skill (Kröger *et al* 2018, Bilbao *et al* 2021). Also, those initial conditions used for the model initialization are based on observation-based products, which may not be of sufficient quality, especially over regions with poor observational coverage. Besides, the drivers for different variables and extreme indices are different, thus contributing to the region-dependent impact of initialization.

The maps of residual correlation for temperature extremes show a different pattern compared to that for TAS, and there is a generally lower added skill for extreme predictions. Still, some regions show a positive impact of initialization, and some of them are similar between TAS and extreme temperature (e.g. southern part of Australia). For example, minimum temperature extremes (TN10p and TNn; figures 2(c) and (f), respectively) show significantly positive values over parts of South America, Africa, and Australia. However, the TN10p index shows a larger fraction of the global region with negative than positive significance, meaning that DCPP capture less variability than HIST. The different patterns of residual correlations for mean and extreme temperature may be due to annual averages being aggregated over the whole year, thus having multiple different weather conditions. However, extremes represent a small number of days when specific weather situations and drivers may occur and play a role, therefore showing a different pattern for the impact of initialization compared to that for the mean. Also, the temporal aggregation to create annual means smooths the time series, removing the high-frequency temporal variations and thus


(a) TAS [fract+ = 23.6%; fract- = 4.3%]

(b) PR [fract+ = 4.7%; fract- = 1.4%]

(c) TN10p [fract+ = 12.9%; fract- = 17.2%]

(d) TX90p [fract+ = 8.1%; fract- = 4.8%]

(e) R95p [fract+ = 4%; fract- = 1.7%]

(f) TNn [fract+ = 5.4%; fract- = 1.8%]

(g) TXx [fract+ = 7.6%; fract- = 2.3%]

(h) Rx5day [fract+ = 1.6%; fract- = 0.7%]

Figure 2. Maps of residual correlation obtained with the DCPP multi-model ensemble with respect to the HIST multi-model ensemble for the forecast years 1–5 (annual means) for the mean variables and extreme indices. The percentage of the global area with statistically significant positive or negative values is shown in the titles. The reference datasets used for the mean near-surface air temperature and PR are the GHCNv4 and the GPCC datasets, respectively. The reference datasets used for the temperature and PR indices are the BEST and REGEN datasets, respectively. Grey colors over land regions correspond to those grid points with missing values in the reference dataset. Crosses indicate that the values are not statistically significant using a two-sided t-test accounting for autocorrelation and controlling the FDR with $\alpha_{FDR} = 0.1$.

removing some noise and making them more predictable than extreme indices based on daily data.

As for PR, the maps of residual correlation for PR extremes are noisy, and the regions with positive significance are mainly restricted over parts of western Africa and South America (figures 2(e) and (h)).

The maps of RPSS show no added skill from initialization for the probabilistic forecasts, with RPSS values being negative or close to zero (figure S4). Besides, the fraction of the global region with significantly negative values is higher than for positive significance. The results obtained with HadEX3 (figures S5 and S6) show that, for minimum temperature extremes, most regions with significant improvement due to initialization coincide among both reference datasets. The results differ more for maximum temperature extremes, particularly for the TX90p index over Asia (higher added skill when using HadEX3 as reference) and TXx index over America (higher added skill when using BEST as reference).

4.3. Multi-model skill for seasonal extreme indices

In addition to the annual extremes, skillful predictions of seasonal extremes are highly important as their impacts on climate-dependent sectors, nature, and society may be more relevant in particular seasons of the year.

As for annual TAS, the DCPP multi-model ensemble shows high skill in predicting TAS in both DJF and JJA, with 92.6% and 97.1% of the global land area being statistically significant, respectively (figures 3(a) and (b)). This fraction is slightly lower than for annual means (99.5%; figure 1(a)). For temperature extremes, the skill patterns differ between seasons. Overall, there is a higher skill and larger areas showing statistically significant skill for temperature extremes during JJA than DJF (figures 3(c)–(j)). For instance, no significance is found around the Mediterranean region for DJF extreme predictions, while most grid points show statistical significance during JJA. These results are consistent with Hanlon *et al* (2013), who also found higher-quality predictions in the Mediterranean region for JJA than for DJF. However, there are also regions where some indices are predicted more skillfully during DJF than during JJA (for example, TN10p over Central Canada and TNn over some parts of South America). Besides, the skill is generally higher for (more moderate) percentile-based than for (more intense) absolute extremes, consistent with Hanlon *et al* (2015).

The skill patterns also differ between seasons for PR. In DJF, positive skill is restricted to Northern Europe and some parts of Central and Eastern Asia, although without significance (0.5% of the global fraction; figure 3(k)). In JJA, the regions showing statistical significance are mainly located over South America and Northern Africa, representing 6.7% of the global region (figure 3(m)). The skill for extreme PR (figures 3(l) and (n)) shows overall similar patterns as those for PR (figures 3(k) and (m)).

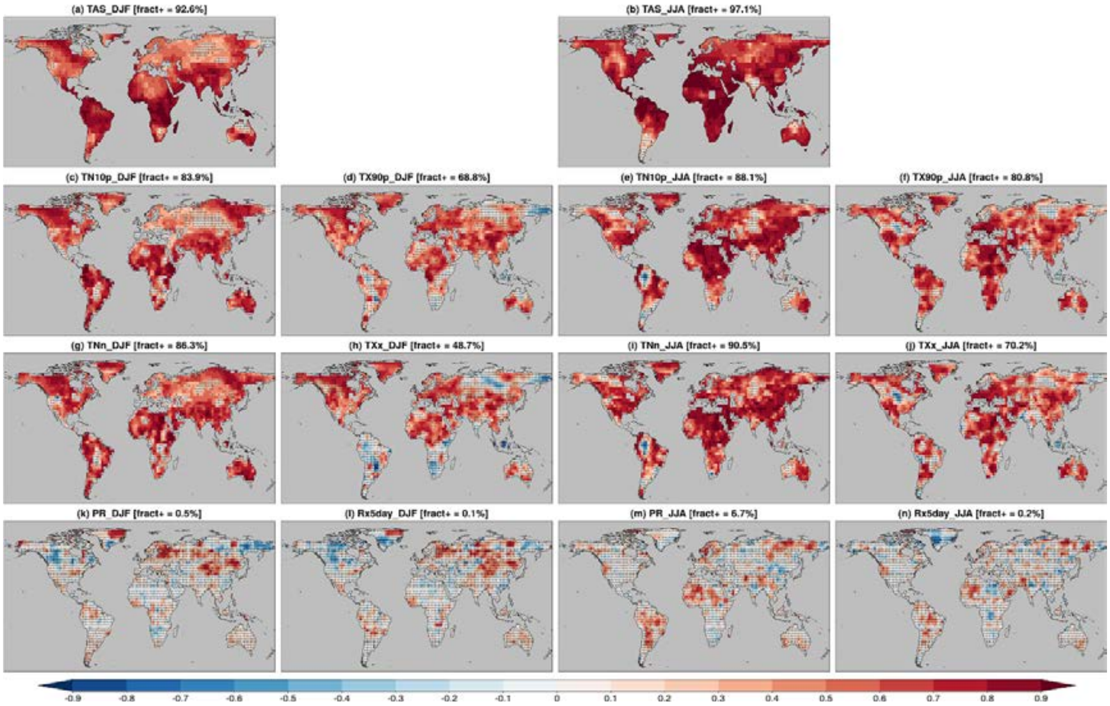

(a) TAS_DJF [fract+ = 92.6%]
(b) TAS_JJA [fract+ = 97.1%]
(c) TN10p_DJF [fract+ = 83.9%]
(d) TX90p_DJF [fract+ = 68.8%]
(e) TN10p_JJA [fract+ = 88.1%]
(f) TX90p_JJA [fract+ = 80.8%]
(g) TNn_DJF [fract+ = 86.3%]
(h) TXx_DJF [fract+ = 48.7%]
(i) TNn_JJA [fract+ = 90.5%]
(j) TXx_JJA [fract+ = 70.2%]
(k) PR_DJF [fract+ = 0.5%]
(l) Rx5day_DJF [fract+ = 0.1%]
(m) PR_JJA [fract+ = 6.7%]
(n) Rx5day_JJA [fract+ = 0.2%]
-0.9
-0.8
-0.7
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

Figure 3. Maps of ACC obtained with the DCPP multi-model ensemble for the forecast years 1–5 (boreal winter and summer means) for the mean variables and extreme indices. The percentage of the global area with statistically significant positive or negative values is shown in the titles. The reference datasets used for the mean near-surface air temperature and PR are the GHCNv4 and the GPCC datasets, respectively. The reference datasets used for the temperature and PR indices are the BEST and REGEN datasets, respectively. Grey colors over land regions correspond to those grid points with missing values in the reference dataset. Crosses indicate that the values are not statistically significant using a one-sided t-test accounting for autocorrelation and controlling the FDR with $\alpha_{FDR} = 0.1$.

The seasonality of the skill may be caused by several factors. For instance, summer PR over Central Europe is more convective (Llasat *et al* 2021), which may decrease the prediction skill because the limited spatial resolution of climate models does not allow for resolving the small-scale processes (Merryfield *et al* 2020). Also, the large-scale drivers and their local response are different across seasons (Müller *et al* 2012), thus affecting the prediction skill. For example, (Palmer *et al* 2008) showed that the skill of seasonal forecasts for winter European heat waves is reduced due to model deficiencies in representing blocking systems. Besides, the spatial distribution of trends also differs (Lee *et al* 2021), as well as the signal-to-noise ratio (Schubert *et al* 2009), which also may contribute to the seasonality of the prediction skill.

The probabilistic evaluation shows a general benefit from using DCPP with respect to the climatological forecast for mean and extreme temperature, as the RPSS is positive and significant over large regions of the globe (figure S7). For PR, although lower RPSS values are found, there are regions where DCPP outperform the climatological forecast. For instance, significantly positive RPSS values are found over several regions of Eurasia for mean and extreme PR during DJF, and over Central Africa during JJA. The results are generally consistent with those obtained with HadEX3 (figures S8 and S9). The highest discrepancies are found for some extreme temperature indices, especially over the Americas.

4.4. Impact of initialization for seasonal extreme indices

The impact of initialization for seasonal TAS shows different patterns during DJF and JJA (figures 4(a) and (b), respectively). For DJF temperature predictions, the greatest added skill is found in Africa, particularly over some Central and Northern regions. Other regions like Southern Australia, Southern Asia and some parts of the Americas also significantly benefit from initialization. Still, negative residual correlation is found over regions like northern Eurasia. During JJA, significant added skill is mainly found over Central Asia, Northeastern Africa, and some parts of the Americas, Australia and Europe. Contrarily, significant skill decrease is found over large regions of North America, Africa and Asia.

The regions where a significant impact of initialization is found are different between the extreme temperature indices (figures 4(c)–(j)) and are also different from those obtained for TAS. The fraction of the globe showing a significantly added skill is higher than the fraction showing a worsening for all the indices considered. Besides, the fraction of the

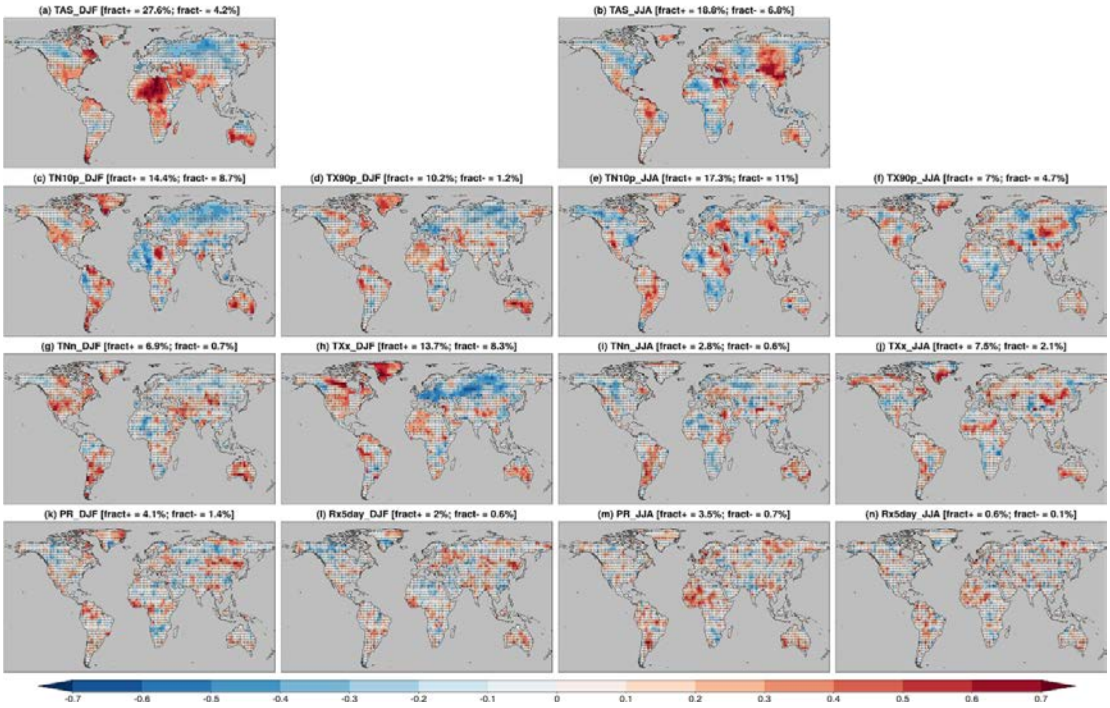

(a) TAS_DJF [fract+ = 27.6%; fract- = 4.2%]
(b) TAS_JJA [fract+ = 18.8%; fract- = 6.8%]
(c) TN10p_DJF [fract+ = 14.4%; fract- = 8.7%]
(d) TX90p_DJF [fract+ = 10.2%; fract- = 1.2%]
(e) TN10p_JJA [fract+ = 17.3%; fract- = 11%]
(f) TX90p_JJA [fract+ = 7%; fract- = 4.7%]
(g) TNn_DJF [fract+ = 6.9%; fract- = 0.7%]
(h) TXx_DJF [fract+ = 13.7%; fract- = 8.3%]
(i) TNn_JJA [fract+ = 2.8%; fract- = 0.6%]
(j) TXx_JJA [fract+ = 7.5%; fract- = 2.1%]
(k) PR_DJF [fract+ = 4.1%; fract- = 1.4%]
(l) Rx5day_DJF [fract+ = 2%; fract- = 0.6%]
(m) PR_JJA [fract+ = 3.5%; fract- = 0.7%]
(n) Rx5day_JJA [fract+ = 0.6%; fract- = 0.1%]
-0.7
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

Figure 4. Maps of residual correlation obtained with the DCPP multi-model ensemble with respect to the HIST multi-model ensemble for the forecast years 1–5 (boreal winter and summer means) for the mean variables and extreme indices. The percentage of the global area with statistically significant positive or negative values is shown in the titles. The reference datasets used for the mean near-surface air temperature and PR are the GHCNv4 and the GPCC datasets, respectively. The reference datasets used for the temperature and PR indices are the BEST and REGEN datasets, respectively. Grey colors over land regions correspond to those grid points with missing values in the reference dataset. Crosses indicate that the values are not statistically significant using a two-sided t-test accounting for autocorrelation and controlling the FDR with $\alpha_{FDR} = 0.1$.

global region showing added skill from initialization is higher for seasonal indices than for annual indices. Still, there is a smaller added skill for all indices than for TAS. Hanlon *et al* (2013) found that initialization did not improve the skill for European summer extremes with the CMIP5 multi-model ensemble. However, they found added skill when assessing the MPI-ESM1.2-LR model individually. This points to the importance of also assessing each forecast system individually, as the multi-model ensemble does not necessarily outperform all single forecast systems (Mishra *et al* 2018, Delgado-Torres *et al* 2022).

For PR, the patterns also differ between seasons (figures 4(k) and (m)), being the one for JJA more similar to that for annual means (figure 2(b)). Besides, the patterns for seasonal extreme indices (figures 4(l) and (m)) are similar to those for seasonal means.

There is no or limited added skill from initialization for the probabilistic forecasts, as the RPSS values are close to zero for the seasonal means and extreme indices of both temperature and PR (figure S10). The results obtained using HadEX3 show similar results (figures S11 and S12, respectively).

5. Summary and conclusion

We have evaluated some aspects of the forecast quality of the CMIP6 decadal forecast systems in predicting TAS, PR, and a set of extreme indices based on daily minimum and maximum temperatures and PR for predictions of the next five years, considering annual, DJF and JJA extremes. The prediction skill of deterministic and probabilistic forecasts has been estimated and compared to that of the HIST to assess the impact of model initialization.

The DCPP multi-model shows high skill in predicting mean and extreme TAS indices computed at annual frequency over most of the globe. The skill is lower and limited to some regions for mean and extreme PR. There is a generally higher skill in predicting the mean variables than the extreme indices. The skill for both extreme TAS and PR is higher for the moderate extremes (TN10p, TX90p and R95p; related to frequency) than for the most extreme extremes (TNn, TXx and Rx5day, related to intensity). The comparison between DCPP and HIST shows a region-dependent impact of initialization on the skill. The added skill due to initialization is higher for the mean variables than for the extreme indices. Besides, such skill differences differ between indices, especially those representing extreme temperature.

For seasonal means and seasonal extreme indices, the DCPP multi-model also shows a generally high skill for mean and extreme TAS, especially during JJA. Similar to annual indices, the skill is higher for the moderate extreme indices than for the most intense extremes for both DJF and JJA. The skill is also more limited for mean and extreme PR than for temperature on seasonal time scales. Still, there is high skill during DJF over Northern Europe and some regions of Asia, and during JJA over some regions of South America, Central Africa and Northeastern Asia. The residual correlations show different patterns for seasonal mean variables and extreme indices compared to those for annual frequency. For temperature, a lower added skill is found for extreme than for mean temperature. In addition, the added skill for extreme temperature depends on the region and season. For PR, the impact of initialization is similar for extreme and mean PR. Comparing seasons, the added skill is more similar between JJA and annual, while it differs more for DJF.

In conclusion, we find that the CMIP6 decadal forecast systems can skillfully predict characteristics of climate extremes, in particular extremely hot and cold temperatures. While the prediction skill for the extremes indices is mostly lower than for annual or seasonal means, these forecast systems still provide useful predictions for the more impact-relevant aspects of climate. However, to exploit all the potential usefulness of decadal predictions, user-oriented indicators could be explored to facilitate their applicability in climate-sensitive sectors, which might be based on variables other than temperature and precipitation, such as wind speed and solar radiation. Besides, the analogous forecast quality assessment should be performed for the particular region, index and forecast period for each user-specific need (Hanlon *et al* 2015, Sgubin *et al* 2021). This systematic evaluation of decadal hindcasts is essential when providing a climate service based on decadal predictions so that the user is informed on the trustworthiness of the forecasts. Also, comparing decadal hindcasts and historical simulations might help climate services providers to select the highest-quality climate information for each particular case.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://esgf-node.llnl.gov/search/cmip6/.

Acknowledgments

This study is a contribution to the Horizon Europe ASPECT project (Grant 101081460). C D T acknowledges financial support from the Spanish Ministry for Science and Innovation (FPI PRE2019-509 08864 financed by MCIN/AEI/10.13039/501100011033 and by FSE invierte en tu futuro). M G D is grateful for support by the AXA Research Fund. We acknowledge the use of the climdex.pcic, startR, s2dv, CSTools, multiApply, and ClimProjDiags R-language-based software packages, all of them available on the Comprehensive R Archive Network (CRAN; https://cran.r-project.org/). The authors thank all the modeling groups for making the simulations available through the Earth System Grid Federation system (ESGF; https://esgf-node.llnl.gov/search/cmip6/).

ORCID iDs

Carlos Delgado-Torres https://orcid.org/0000-0003-1737-4212
Markus G Donat https://orcid.org/0000-0002-0608-7288
Albert Soret https://orcid.org/0000-0002-1962-2972
Nube González-Reviriego https://orcid.org/0000-0002-5919-6701
Pierre-Antoine Bretonnière https://orcid.org/0000-0002-3066-6685
An-Chi Ho https://orcid.org/0000-0002-4182-5258
Núria Pérez-Zanón https://orcid.org/0000-0001-8568-3071
Margarida Samsó Cabré https://orcid.org/0000-0003-2868-2755
Francisco J Doblas-Reyes https://orcid.org/0000-0002-6622-4280

References

Alexander L V, Bador M, Roca R, Contractor S, Donat M G and Nguyen P L 2020 *Environ. Res. Lett.* **15** 055002

Alexander L V, Uotila P and Nicholls N 2009 *J. Geophys. Res.: Atmos.* **114** 18116

Bhend J, Mahlstein I and Liniger M A 2017 *Q. J. R. Meteorol. Soc.* **143** 184–94

Bilbao R *et al* 2021 *Earth Syst. Dyn.* **12** 173–96

Boer G J *et al* 2016 *Geosci. Model Dev.* **9** 3751–77

Borchert L F, Menary M B, Swingedouw D, Sgubin G, Hermanson L and Mignot J 2021 *Geophys. Res. Lett.* **48** e2020GL091307

Bronaugh D 2020 *Climdex.pcic: Pacific Climate Impacts Consortium (PCIC) Implementation of Climdex Routines* R package version 1.1-11 (available at: https://CRAN.R-project.org/package=climdex.pcic)

Contractor S, Donat M G, Alexander L V, Ziese M, Meyer-Christoffer A, Schneider U, Rustemeier E, Becker A, Durre I and Vose R S 2020 *Hydrol. Earth System Sci.* **24** 919–43

Curtis S, Fair A, Wistow J, Val D V and Oven K 2017 *Environ. Health: A Glob. Access Sci. Source* **16** 23–32

Delgado-Torres C *et al* 2022 *J. Clim.* **35** 4363–82

Delsole T and Tippett M K 2016 *Mon. Weather Rev.* **144** 615–26

Doblas-Reyes F J, Andreu-Burillo I, Chikamoto Y, García-Serrano J, Guemas V, Kimoto M, Mochizuki T, Rodrigues L R L and Oldenborgh G J V 2013 *Nat. Commun.* **4** 1715

Dunn R J *et al* 2020 *J. Geophys. Res.: Atmos.* **125** e2019JD032263

Eade R, Hamilton E, Smith D M, Graham R J and Scaife A A 2012 *J. Geophys. Res.: Atmos.* **117** 21110

Eyring V, Bony S, Meehl G A, Senior C A, Stevens B, Stouffer R J and Taylor K E 2016 *Geosci. Model Dev.* **9** 1937–58

Hanlon H M, Hegerl G C, Tett S F and Smith D M 2013 *J. Clim.* **26** 3728–44

Hanlon H M, Hegerl G C, Tett S F and Smith D M 2015 *Clim. Change* **132** 61–76

Hanlon H M, Morak S and Hegerl G C 2013 *J. Geophys. Res.: Atmos.* **118** 9631–41

Kröger J *et al* 2018 *Clim. Dyn.* **51** 2593–608
Kushnir Y *et al* 2019 *Nat. Clim. Change* **9** 94–101
Lee J Y *et al* 2021 *Climate Change 2021: The Physical Science Basis (Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change)* ed V Masson-Delmotte (Cambridge: Cambridge University Press) sec 4
Liu Y, Donat M G, Rust H W, Alexander L V and England M H 2019 *Clim. Dyn.* **53** 3711–29
Llasat M C, del Moral A, Cortès M and Rigo T 2021 *Atmos. Res.* **257** 105581
Mahlstein I, Spirig C, Liniger M A and Appenzeller C 2015 *J. Geophys. Res.: Atmos.* **120** 2808–18
Meehl G A *et al* 2009 *Bull. Am. Meteorol. Soc.* **90** 1467–86
Meehl G A *et al* 2021 *Nat. Rev. Earth Environ.* **2** 340–57
Menne M J, Williams C N, Gleason B E, Rennie J J and Lawrimore J H 2018 *J. Clim.* **31** 9835–54
Merryfield W J *et al* 2020 *Bull. Am. Meteorol. Soc.* **101** E869–96
Mishra N, Prodhomme C and Guemas V 2018 *Clim. Dyn.* **52** 4207–25
Moemken J, Feldmann H, Pinto J G, Buldmann B, Laube N, Kadow C, Paxian A, Tiedje B, Kottmeier C and Marotzke J 2021 *Int. J. Climatol.* **41** E1944–58
Müller W A, Baehr J, Haak H, Jungclaus J H, Kröger J, Matei D, Notz D, Pohlmann H, von Storch J S and Marotzke J 2012 *Geophys. Res. Lett.* **39** L22707
Palmer T N, Doblas-Reyes F J, Weisheimer A and Rodwell M J 2008 *Bull. Amer. Meteor. Soc.* **89** 459–70
Scaife A A and Smith D 2018 *npj Clim. Atmos. Sci.* **1** 1–8
Schneider U, Becker A, Finger P, Rustemeier E and Ziese M 2020 Global Precipitation Climatology Centre (GPCC, Deutscher Wetterdienst) (available at: http://gpcc.dwd.de/)
Schubert S *et al* 2009 *J. Clim.* **22** 5251–72
Seneviratne S *et al* 2021 *Climate Change 2021: The Physical Science Basis (Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climatechange)* ed V Masson-Delmotte (Cambridge: Cambridge University Press) sec 11
Sgubin G, Swingedouw D, Borchert L F, Menary M B, Noël T, Loukos H and Mignot J 2021 *Clim. Dyn.* **57** 3245–63
Sillmann J, Kharin V V, Zhang X, Zwiers F W and Bronaugh D 2013 *J. Geophys. Res.: Atmos.* **118** 1716–33
Sillmann J, Thorarinsdottir T, Keenlyside N, Schaller N, Alexander L V, Hegerl G, Seneviratne S I, Vautard R, Zhang X and Zwiers F W 2017 *Weather Clim. Extremes* **18** 65–74
Smith D M *et al* 2019 *npj Clim. Atmos. Sci.* **2** 1–10
Smith D M, Cusack S, Colman A W, Folland C K, Harris G R and Murphy J M 2007 *Science* **317** 796–9
Smith D M, Eade R, Dunstone N J, Fereday D, Murphy J M, Pohlmann H and Scaife A A 2010 *Nat. Geosci.* **3** 846–9
Taylor K E, Stouffer R J and Meehl G A 2012 *Bull. Am. Meteorol. Soc.* **93** 485–98
van Ulden A P and van Oldenborgh G J 2006 *Atmos. Chem. Phys.* **6** 863–81
Wilks D S 2011 *Int. Geophys.* **100** 301–94
Wilks D S 2016 *Bull. Am. Meteorol. Soc.* **97** 2263–73
Zhang X, Alexander L, Hegerl G C, Jones P, Tank A K, Peterson T C, Trewin B and Zwiers F W 2011 *Wiley Interdiscip. Rev. Clim. Change* **2** 851–70
Zwiers F W and von Storch H 1995 *J. Clim.* **8** 336–51

Chapter 6

Discussion and applications

Decadal climate predictions have been shown to provide high-quality climate information for some variables and regions at annual to decadal time scales, therefore enabling their applicability to support decision-making in climate-dependent sectors and adapt to the consequences of climate variability and change. In addition to the results shown in the research articles published in peer-reviewed scientific journals, several additional studies and research questions have been addressed during this Ph.D., developed as additional research within the peer-reviewed articles presented in Chapters 3, 4 and 5, and Spanish and European projects. Such results can be accessed through an R Shiny App: https://earth.bsc.es/shiny/cdelgado/. This Shiny App and those mentioned below are restricted for privacy reasons. Please send an email to carlos.delgado@bsc.es to get the credentials.

In this Ph.D. thesis, annual to decadal forecasts of essential climate variables and modes of variability (Chapter 3; Delgado-Torres *et al.*, 2022a), spatial patterns and frequencies of occurrence of the Euro-Atlantic weather regimes (Chapter 4; Delgado-Torres *et al.*, 2022b), indices that account for the frequency and intensity of extreme climate events (Chapter 5; Delgado-Torres *et al.*, 2023), and the SPEI for drought conditions estimation (below in this Chapter) have been evaluated motivated by their potential applicability for climate services on climate-sensitive sectors. Nevertheless, these variables and indices might not be the most suitable for specific decision-making for all users. Thus, it is necessary to discuss with users what and how they make decisions to produce tailored indicators. Also, it should be noted that the results obtained may differ for other variables, indices, regions and forecast periods. Thus, such results should not be extrapolated to other cases, and a careful evaluation should be carried out for each particular purpose.

Several metrics have been selected to assess the most general aspects of the forecast quality. These metrics include, for instance, the Anomaly Correlation Coefficient (ACC; Wilks, 2011) and Root Mean Squared Error (RMSE; Wilks, 2011) to evaluate the deterministic forecasts, and the Brier Score (BS; Wilks, 2011), Ranked Probability Score (RPS; Wilks, 2011), Continuous Ranked Probability Score (CRPS; Wilks, 2011), Relative Operating Characteristic (ROC; Kharin & Zwiers, 2003) and Spread-to-Error ratio (Hopson, 2014) to evaluate the probabilistic forecasts. Nonetheless, the type of final product that users request may be different. For example, some users may request a product based only on the ensemble mean, the ensemble mean plus the ensemble spread, a probabilistic forecast based on tercile/quintile categories or the full probability distribution function. Therefore, such a specific product should be evaluated accordingly using the skill metric that assesses the particular aspect of the predictions that impact their societal or economic decisions the most.

In addition to selecting the most suitable skill metric, the reference forecast for skill score calculation should also consider what users have been using before receiving climate information based on numerical or statistical climate models. In this Ph.D. thesis, the climatological forecast and historical forcing simulations have been used as reference forecasts to estimate the benefit of using decadal predictions. However, different users may have been using other classical forecasts, such as predictions based on persistence (i.e. based on the climate conditions during a recent period), a combination of persistence and climatology (Murphy, 1992), or predictions based on climate trends.

The post-processing techniques (such as multi-model combination, calibration and downscaling) should also be selected accounting for the final product to deliver the most useful climate information. Within this thesis, several multi-model approaches have been compared, finding non-significant differences between them. These multi-model approaches used the information from all available decadal prediction systems without weighting them according to their skill. Using weight-based multi-model approaches could increase the forecast quality as skillful models would have more importance within the multi-model ensemble. However, Mishra *et al.* (2018) found no benefit of weighting forecast systems to increase the quality of seasonal predictions. The choice of the calibration method should also account for the specific aspect of the forecast quality that aims to be improved, as each technique corrects a certain statistical property of the predictions but may degrade others. In addition, applying calibration may decrease the quality due to the mere fact of performing it in cross-validation mode. However, the price to pay for calibration is rewarded as the predictions need to have the observed statistical properties and increase their reliability (i.e. agreement between forecast probability and mean observed frequency; Murphy, 1993) for them to be usable for users and decision-makers.

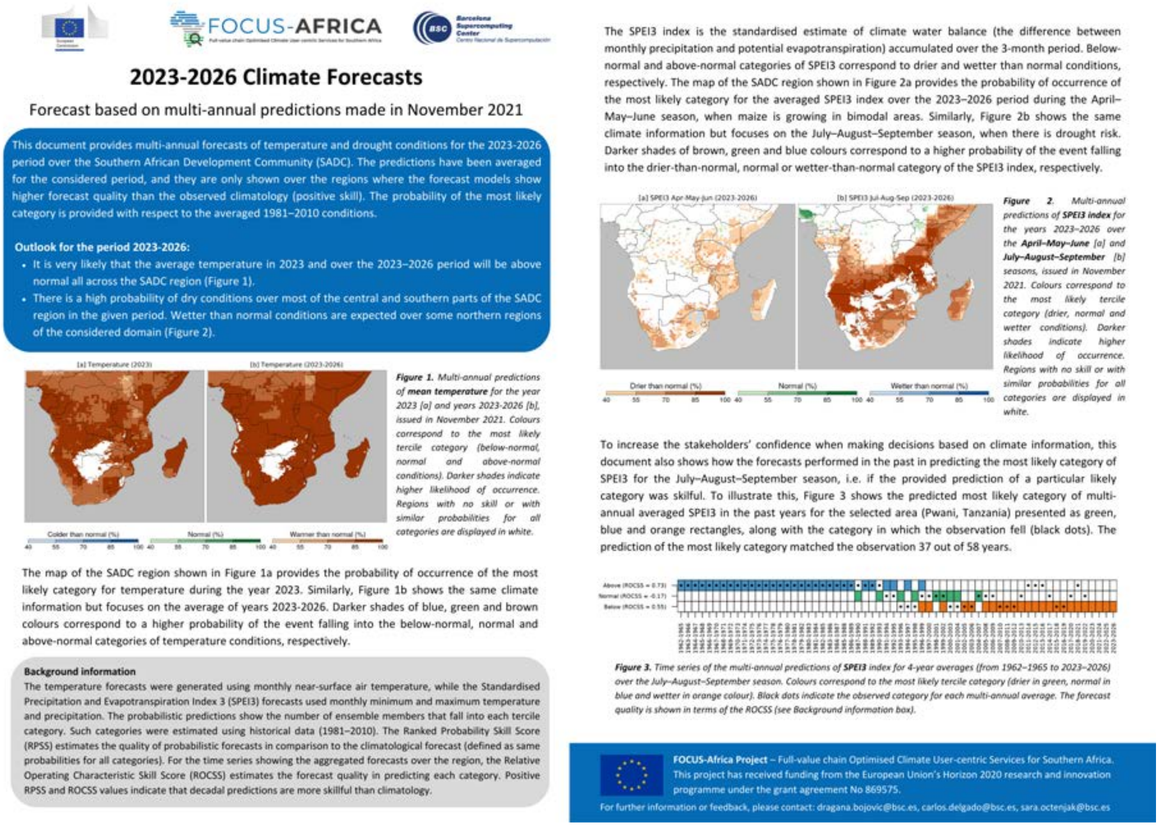
In comparison to uninitialised historical forcing simulations, model initialisation is expected to increase the quality of decadal predictions as climate models incorporate information on current conditions, thus phasing the simulations with observations and narrowing the uncertainty of the forecasts. However, the impact of initialisation is not always positive, as initialisation shocks and errors in initial conditions can decrease the quality of the predictions. Therefore, to provide the best climate information, it is necessary to perform a forecast quality assessment for the variables, indicators, forecast periods and regions for the particular climate service provision. This is supported by the skill differences and distinct impacts of initialisation found during the systematic forecast evaluations and comparison between initialised decadal predictions and uninitialised historical simulations.

One of the major barriers to developing climate services is the availability of near real-time predictions. While there is currently decadal hindcast data (which corresponds to the Component A of the CMIP6/DCPP) available from 13 forecast systems on the ESGF portals, decadal forecasts data (Component B of the CMIP6/DCPP) is only available from a small subset of forecast systems. For instance, there are near real-time predictions of mean temperature and precipitation from only 6 forecast systems. On top of this, the number of systems also depends on the variable, as there is more data available at monthly aggregations than at daily frequency, which poses an extra problem for the development of climate services based on extreme indices computed, e.g. from daily temperature and precipitation. Another obstacle for the climate services provision is the timeliness of the data delivery, especially if the users are interested in forecast periods that include the first forecast year.

Climate services based on decadal predictions are in a early state, as only a few studies based on case studies have applied climate information extracted from decadal predictions to practical application and illustration on how predictions at annual to decadal time scales can support the decision-making process (e.g. Paxian *et al.*, 2019; Solaraju-Murali, 2023). In this thesis, the potential applicability of decadal predictions has also been tested for case studies on the food-security sector over the Southern African Development Community (SADC) region within the framework of the FOCUS-Africa project (https://focus-africaproject.eu/).

In this context, two forecast products in the form of two-pager and four-pager documents were developed (Figures 6.1 and 6.2). They included downscaled forecasts of temperature (both mean and extreme), precipitation and drought conditions (using the SPEI with different accumulation periods) for the 2022-2026 period over the SADC and Tanzania regions using decadal predictions produced at the end of 2021. An R Shiny App was created to share all the skill estimates and forecast products with the stakeholders: https://earth.bsc.es/shiny/cdelgado_FOCUS-Africa-casestudy/. This work is still in progress and the users' feedback is being considered in order to adapt and tailor such products to the users' needs. For instance, in the first version of the products, a no-skill mask was applied over those regions where the multi-model ensemble does not show an added value in comparison to the climatological forecast. Users expressed their will to have another forecast that could be based, for example, on climatology or persistence. This adaptation is not trivial, as the probabilistic version of the climatological forecast is defined as the same probability of occurrence for all categories (33.3% in case of tercile categories), thus having no most likely tercile. Another related issue is the choice of tercile categories for the probabilistic forecast. Dividing the distribution function in tercile is the typical choice in the literature, but it is not necessarily the most useful choice for users. Other options include, for example, using two or five categories, or the full probability distribution function.

On the second interaction with stakeholders, they showed interests on having two additional indices related the extreme precipitations. These indices are the Consecutive Dry Days and Consecutive Wet Days (CDD and CWD, respectively; Zhang *et al.*, 2011). Regarding the forecast system used to create the climate information, the stakeholders were in favour of using the most skilful source of information for each case, which can be based on a forecast system, multi-model ensemble, climatology or persistence. This implies to select a different system for each variable, region, and forecast period, as suggested in Delgado-Torres *et al.* (2022a) (Chapter 3 of this thesis). Also, there was an agreement on the use of both tercile and quintile categories, as it is the setting for


FOCUS-AFRICA
Full-value chain Optimised Climate User-centric Services for Southern Africa

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

2023-2026 Climate Forecasts

Forecast based on multi-annual predictions made in November 2021

This document provides multi-annual forecasts of temperature and drought conditions for the 2023-2026 period over the Southern African Development Community (SADC). The predictions have been averaged for the considered period, and they are only shown over the regions where the forecast models show higher forecast quality than the observed climatology (positive skill). The probability of the most likely category is provided with respect to the averaged 1981–2010 conditions.

Outlook for the period 2023-2026:

- It is very likely that the average temperature in 2023 and over the 2023–2026 period will be above normal all across the SADC region (Figure 1).
- There is a high probability of dry conditions over most of the central and southern parts of the SADC region in the given period. Wetter than normal conditions are expected over some northern regions of the considered domain (Figure 2).

*Figure 1. Multi-annual predictions of **mean temperature** for the year 2023 [a] and years 2023-2026 [b], issued in November 2021. Colours correspond to the most likely tercile category (below-normal, normal and above-normal conditions). Darker shades indicate higher likelihood of occurrence. Regions with no skill or with similar probabilities for all categories are displayed in white.*

The map of the SADC region shown in Figure 1a provides the probability of occurrence of the most likely category for temperature during the year 2023. Similarly, Figure 1b shows the same climate information but focuses on the average of years 2023-2026. Darker shades of blue, green and brown colours correspond to a higher probability of the event falling into the below-normal, normal and above-normal categories of temperature conditions, respectively.

Background information
The temperature forecasts were generated using monthly near-surface air temperature, while the Standardised Precipitation and Evapotranspiration Index 3 (SPEI3) forecasts used monthly minimum and maximum temperature and precipitation. The probabilistic predictions show the number of ensemble members that fall into each tercile category. Such categories were estimated using historical data (1981–2010). The Ranked Probability Skill Score (RPSS) estimates the quality of probabilistic forecasts in comparison to the climatological forecast (defined as same probabilities for all categories). For the time series showing the aggregated forecasts over the region, the Relative Operating Characteristic Skill Score (ROCSS) estimates the forecast quality in predicting each category. Positive RPSS and ROCSS values indicate that decadal predictions are more skillful than climatology.

The SPEI3 index is the standardised estimate of climate water balance (the difference between monthly precipitation and potential evapotranspiration) accumulated over the 3-month period. Below-normal and above-normal categories of SPEI3 correspond to drier and wetter than normal conditions, respectively. The map of the SADC region shown in Figure 2a provides the probability of occurrence of the most likely category for the averaged SPEI3 index over the 2023–2026 period during the April–May–June season, when maize is growing in bimodal areas. Similarly, Figure 2b shows the same climate information but focuses on the July–August–September season, when there is drought risk. Darker shades of brown, green and blue colours correspond to a higher probability of the event falling into the drier-than-normal, normal or wetter-than-normal category of the SPEI3 index, respectively.

*Figure 2. Multi-annual predictions of **SPEI3 index** for the years 2023–2026 over the **April–May–June** [a] and **July–August–September** [b] seasons, issued in November 2021. Colours correspond to the most likely tercile category (drier, normal and wetter conditions). Darker shades indicate higher likelihood of occurrence. Regions with no skill or with similar probabilities for all categories are displayed in white.*

To increase the stakeholders' confidence when making decisions based on climate information, this document also shows how the forecasts performed in the past in predicting the most likely category of SPEI3 for the July–August–September season, i.e. if the provided prediction of a particular likely category was skilful. To illustrate this, Figure 3 shows the predicted most likely category of multi-annual averaged SPEI3 in the past years for the selected area (Pwani, Tanzania) presented as green, blue and orange rectangles, along with the category in which the observation fell (black dots). The prediction of the most likely category matched the observation 37 out of 58 years.

*Figure 3. Time series of the multi-annual predictions of **SPEI3** index for 4-year averages (from 1962–1965 to 2023–2026) over the July–August–September season. Colours correspond to the most likely tercile category (drier in green, normal in blue and wetter in orange colour). Black dots indicate the observed category for each multi-annual average. The forecast quality is shown in terms of the ROCSS (see Background information box).*

FOCUS-Africa Project – Full-value chain Optimised Climate User-centric Services for Southern Africa. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 869575.
For further information or feedback, please contact: dragana.bojovic@bsc.es, carlos.delgado@bsc.es, sara.octenjak@bsc.es

Figure 6.1: Two-pager forecast product developed within the framework of the FOCUS-Africa project. The document provides predictions of temperature and drought conditions over the SADC region for the 2023-2026 period.

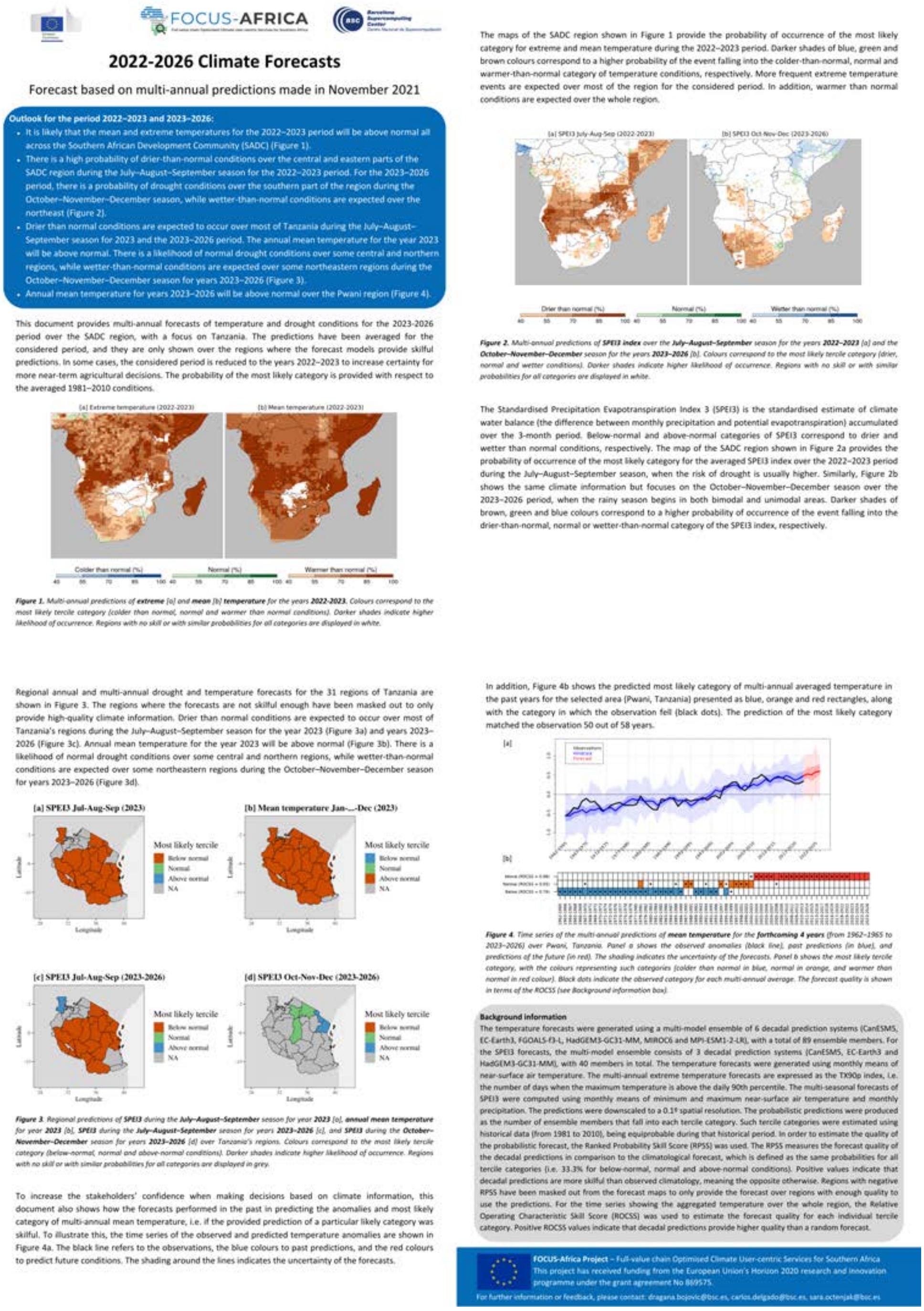
the seasonal predictions within the same project, and the users also requested the use of the same colour palettes as for seasonal predictions.

Another aspect of the products to be improved is the distribution of the information along the documents, as the users requested having only one variable per page to clearly understand what is being shown. They also asked for changing the baseline period used to compute the anomalies and probabilistic categories, replacing the original 1981-2010 period with the 1991-2020 period. In general, they showed satisfaction with the forecast products and said that, without previous contact with climate information at annual to decadal time scales, they see potential for its use for decisions that need a long margin to be implemented. We aim to publish a research article, which is currently in preparation, presenting and describing the entire co-production process as well as how the users' feedback has been taken into account to improve the forecast products.

Another application of climate information at annual to decadal time scales was tested with Decathlon, a multinational sporting goods retailer and manufacturer based in France (https://www.decathlon.com/). Decathlon was interested in predictions of temperature and precipitation for the next few years. This was motivated by the relationship of these climate variables with the cotton production, as variations in its price are modulated by cotton availability and production. Decathlon was interested in predictions for different periods (e.g. forecast year 1, 2, 2-5 and 6-10) over seven different regions where they purchase cotton. The growing season for the cotton depends on the region. Thus, the MIRCA2000 dataset (Portmann *et al.*, 2010, which provides the crop months globally) has been used to select the predictions for specific crop months for each location where the cotton is produced (Figure 6.3). This includes both irrigated and rainfed crops.

During the first phase of the contract, the work focused on the forecast quality assessment of predictions for the settings mentioned above. The first phase of the contract consisted of performing a systematic forecast quality assessment using a set of different skill metrics. For instance, Figure 6.4 shows such metrics for multi-model predictions of precipitation for the forecast crop seasons 2-5. The results were presented in meetings with the company as well as into a document summarising the main results. In addition, an R Shiny App was developed to facilitate the access and visualisation of the results: https://earth.bsc.es/shiny/cdelgado_Decathlon/. The second phase of the contract consisted of the provision of the forecasts through a document and a shiny app as well as the presentation of such forecasts in a meeting. These forecasts include both deterministic (based on the ensemble mean) and probabilistic products (based on both the most likely category and the likelihood for each category). As a forecast example, Figure 6.5 shows the probability of the most likely tercile forecasted for the 2023-2026 period, selecting only those specific crop months for each location.

Additionally, Decathlon was also provided with regional deterministic and probabilistic forecasts for the areas where they purchase cotton. For instance, Figure 6.6 shows the predicted probability for each tercile category for precipitation over Brazil for the forecast years 2-5. In the same figure, dots indicate the observed category in past years to increase the users' confidence.


FOCUS-AFRICA
Barcelona Supercomputing Center
Centro Nacional de Supercomputación
[a] Extreme temperature (2022-2023)
[b] Mean temperature (2022-2023)
Colder than normal (%)
Normal (%)
Warmer than normal (%)
40
55
70
85
100
[a] SPEI3 July-Aug-Sep (2022-2023)
[b] SPEI3 Oct-Nov-Dec (2023-2026)
Drier than normal (%)
Normal (%)
Wetter than normal (%)
[a] SPEI3 Jul-Aug-Sep (2023)
[b] Mean temperature Jan-...-Dec (2023)
[c] SPEI3 Jul-Aug-Sep (2023-2026)
[d] SPEI3 Oct-Nov-Dec (2023-2026)
Most likely tercile
Below normal
Normal
Above normal
NA
Latitude
Longitude
[a]
Observations
Hindcast
Forecast
[b]

2022-2026 Climate Forecasts

Forecast based on multi-annual predictions made in November 2021

Outlook for the period 2022–2023 and 2023–2026:

- It is likely that the mean and extreme temperatures for the 2022–2023 period will be above normal all across the Southern African Development Community (SADC) (Figure 1).
- There is a high probability of drier-than-normal conditions over the central and eastern parts of the SADC region during the July–August–September season for the 2022–2023 period. For the 2023–2026 period, there is a probability of drought conditions over the southern part of the region during the October–November–December season, while wetter-than-normal conditions are expected over the northeast (Figure 2).
- Drier than normal conditions are expected to occur over most of Tanzania during the July–August–September season for 2023 and the 2023–2026 period. The annual mean temperature for the year 2023 will be above normal. There is a likelihood of normal drought conditions over some central and northern regions, while wetter-than-normal conditions are expected over some northeastern regions during the October–November–December season for years 2023–2026 (Figure 3).
- Annual mean temperature for years 2023–2026 will be above normal over the Pwani region (Figure 4).

This document provides multi-annual forecasts of temperature and drought conditions for the 2023-2026 period over the SADC region, with a focus on Tanzania. The predictions have been averaged for the considered period, and they are only shown over the regions where the forecast models provide skilful predictions. In some cases, the considered period is reduced to the years 2022–2023 to increase certainty for more near-term agricultural decisions. The probability of the most likely category is provided with respect to the averaged 1981–2010 conditions.

Figure 1. *Multi-annual predictions of **extreme** [a] and **mean** [b] **temperature** for the years **2022-2023**. Colours correspond to the most likely tercile category (colder than normal, normal and warmer than normal conditions). Darker shades indicate higher likelihood of occurrence. Regions with no skill or with similar probabilities for all categories are displayed in white.*

The maps of the SADC region shown in Figure 1 provide the probability of occurrence of the most likely category for extreme and mean temperature during the 2022–2023 period. Darker shades of blue, green and brown colours correspond to a higher probability of the event falling into the colder-than-normal, normal and warmer-than-normal category of temperature conditions, respectively. More frequent extreme temperature events are expected over most of the region for the considered period. In addition, warmer than normal conditions are expected over the whole region.

Figure 2. *Multi-annual predictions of **SPEI3 index** over the **July–August–September** season for the years **2022–2023** [a] and the **October–November–December** season for the years **2023–2026** [b]. Colours correspond to the most likely tercile category (drier, normal and wetter conditions). Darker shades indicate higher likelihood of occurrence. Regions with no skill or with similar probabilities for all categories are displayed in white.*

The Standardised Precipitation Evapotranspiration Index 3 (SPEI3) is the standardised estimate of climate water balance (the difference between monthly precipitation and potential evapotranspiration) accumulated over the 3-month period. Below-normal and above-normal categories of SPEI3 correspond to drier and wetter than normal conditions, respectively. The map of the SADC region shown in Figure 2a provides the probability of occurrence of the most likely category for the averaged SPEI3 index over the 2022–2023 period during the July–August–September season, when the risk of drought is usually higher. Similarly, Figure 2b shows the same climate information but focuses on the October–November–December season over the 2023–2026 period, when the rainy season begins in both bimodal and unimodal areas. Darker shades of brown, green and blue colours correspond to a higher probability of occurrence of the event falling into the drier-than-normal, normal or wetter-than-normal category of the SPEI3 index, respectively.

Regional annual and multi-annual drought and temperature forecasts for the 31 regions of Tanzania are shown in Figure 3. The regions where the forecasts are not skilful enough have been masked out to only provide high-quality climate information. Drier than normal conditions are expected to occur over most of Tanzania's regions during the July–August–September season for the year 2023 (Figure 3a) and years 2023–2026 (Figure 3c). Annual mean temperature for the year 2023 will be above normal (Figure 3b). There is a likelihood of normal drought conditions over some central and northern regions, while wetter-than-normal conditions are expected over some northeastern regions during the October–November–December season for years 2023–2026 (Figure 3d).

Figure 3. *Regional predictions of **SPEI3** during the **July–August–September** season for year **2023** [a], **annual mean temperature** for year **2023** [b], **SPEI3** during the **July–August–September** season for years **2023–2026** [c], and **SPEI3** during the **October–November–December** season for years **2023–2026** [d] over Tanzania's regions. Colours correspond to the most likely tercile category (below-normal, normal and above-normal conditions). Darker shades indicate higher likelihood of occurrence. Regions with no skill or with similar probabilities for all categories are displayd in grey.*

To increase the stakeholders' confidence when making decisions based on climate information, this document also shows how the forecasts performed in the past in predicting the anomalies and most likely category of multi-annual mean temperature, i.e. if the provided prediction of a particular likely category was skilful. To illustrate this, the time series of the observed and predicted temperature anomalies are shown in Figure 4a. The black line refers to the observations, the blue colours to past predictions, and the red colours to predict future conditions. The shading around the lines indicates the uncertainty of the forecasts.

In addition, Figure 4b shows the predicted most likely category of multi-annual averaged temperature in the past years for the selected area (Pwani, Tanzania) presented as blue, orange and red rectangles, along with the category in which the observation fell (black dots). The prediction of the most likely category matched the observation 50 out of 58 years.

Figure 4. *Time series of the multi-annual predictions of **mean temperature** for the **forthcoming 4 years** (from 1962–1965 to 2023–2026) over Pwani, Tanzania. Panel a shows the observed anomalies (black line), past predictions (in blue), and predictions of the future (in red). The shading indicates the uncertainty of the forecasts. Panel b shows the most likely tercile category, with the colours representing such categories (colder than normal in blue, normal in orange, and warmer than normal in red colour). Black dots indicate the observed category for each multi-annual average. The forecast quality is shown in terms of the ROCSS (see Background information box).*

Background information

The temperature forecasts were generated using a multi-model ensemble of 6 decadal prediction systems (CanESM5, EC-Earth3, FGOALS-f3-L, HadGEM3-GC31-MM, MIROC6 and MPI-ESM1-2-LR), with a total of 89 ensemble members. For the SPEI3 forecasts, the multi-model ensemble consists of 3 decadal prediction systems (CanESM5, EC-Earth3 and HadGEM3-GC31-MM), with 40 members in total. The temperature forecasts were generated using monthly means of near-surface air temperature. The multi-annual extreme temperature forecasts are expressed as the TX90p index, i.e. the number of days when the maximum temperature is above the daily 90th percentile. The multi-seasonal forecasts of SPEI3 were computed using monthly means of minimum and maximum near-surface air temperature and monthly precipitation. The predictions were downscaled to a 0.1º spatial resolution. The probabilistic predictions were produced as the number of ensemble members that fall into each tercile category. Such tercile categories were estimated using historical data (from 1981 to 2010), being equiprobable during that historical period. In order to estimate the quality of the probabilistic forecast, the Ranked Probability Skill Score (RPSS) was used. The RPSS measures the forecast quality of the decadal predictions in comparison to the climatological forecast, which is defined as the same probabilities for all tercile categories (i.e. 33.3% for below-normal, normal and above-normal conditions). Positive values indicate that decadal predictions are more skilful than observed climatology, meaning the opposite otherwise. Regions with negative RPSS have been masked out from the forecast maps to only provide the forecast over regions with enough quality to use the predictions. For the time series showing the aggregated temperature over the whole region, the Relative Operating Characteristic Skill Score (ROCSS) was used to estimate the forecast quality for each individual tercile category. Positive ROCSS values indicate that decadal predictions provide higher quality than a random forecast.

FOCUS-Africa Project – Full-value chain Optimised Climate User-centric Services for Southern Africa
This project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 869575.
For further information or feedback, please contact: dragana.bojovic@bsc.es, carlos.delgado@bsc.es, sara.octenjak@bsc.es

Figure 6.2: Four-pager forecast product developed within the framework of the FOCUS-Africa project. The document provides predictions of mean and extreme temperature and drought conditions over the SADC and Tanzania regions for the 2022-2026 period.

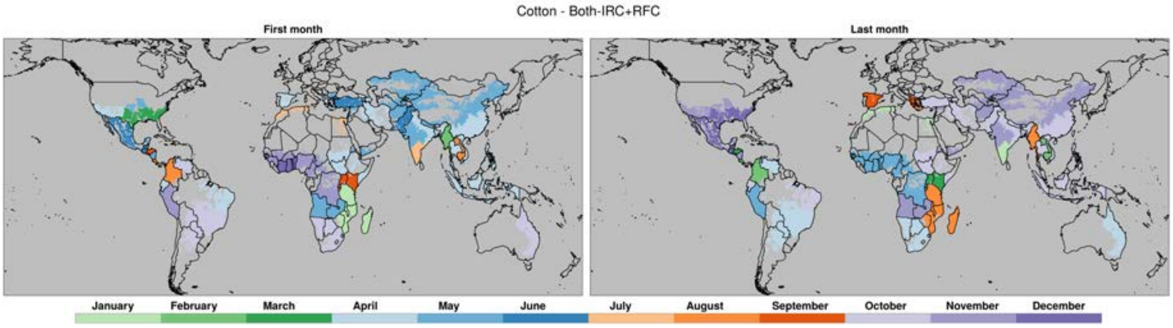


Figure 6.3: First and last months for the cotton crop extracted from the MIRCA2000 dataset for both locations with irrigated and rainfed cotton crop.

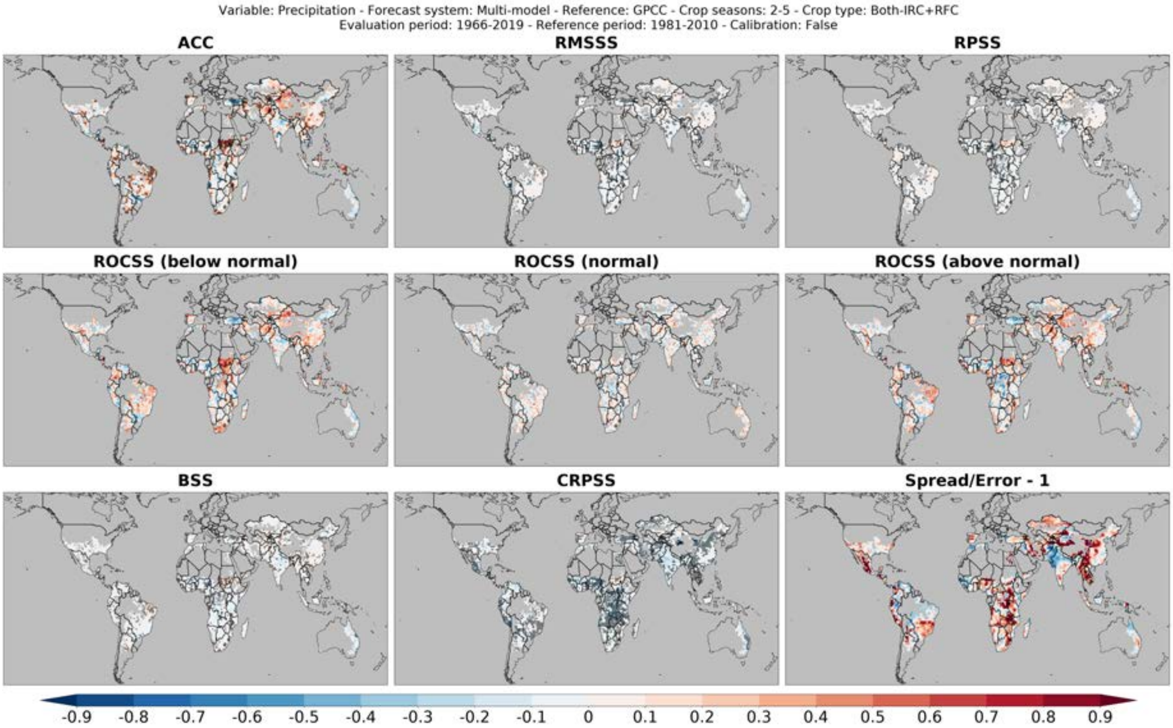

Variable: Precipitation - Forecast system: Multi-model - Reference: GPCC - Crop seasons: 2-5 - Crop type: Both-IRC+RFC
Evaluation period: 1966-2019 - Reference period: 1981-2010 - Calibration: False
ACC
RMSSS
RPSS
ROCSS (below normal)
ROCSS (normal)
ROCSS (above normal)
BSS
CRPSS
Spread/Error - 1
-0.9
-0.8
-0.7
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

Figure 6.4: Forecast quality as measured with different skill metrics obtained with the multi-model ensemble for predictions of precipitation for forecast crop seasons 2-5 over the global cotton crop regions.

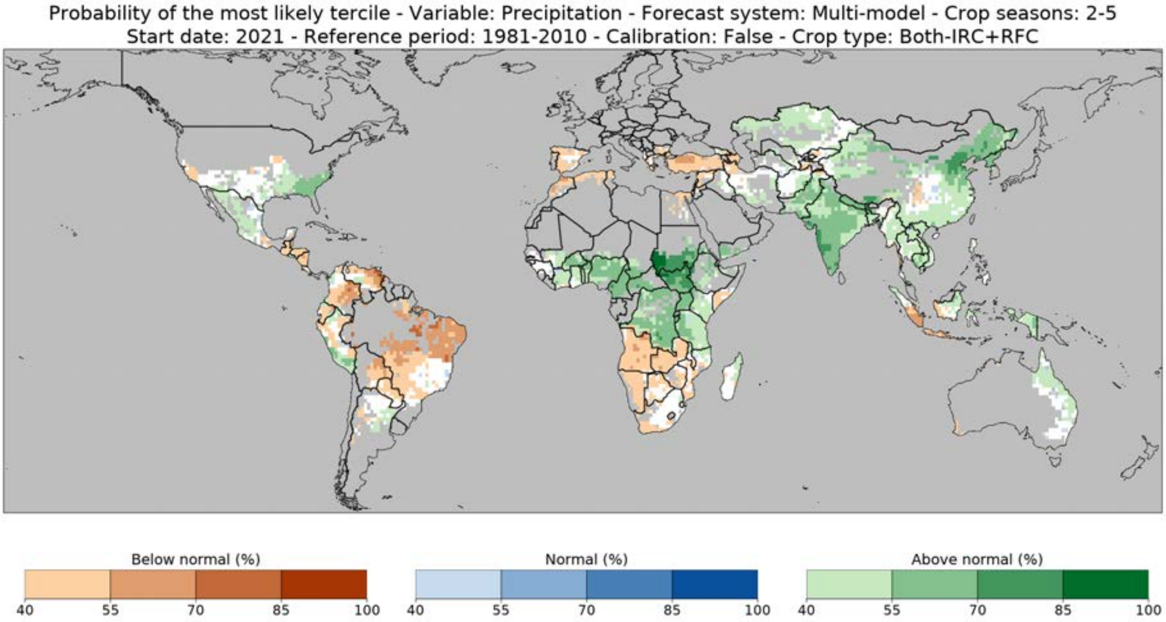

Probability of the most likely tercile - Variable: Precipitation - Forecast system: Multi-model - Crop seasons: 2-5
Start date: 2021 - Reference period: 1981-2010 - Calibration: False - Crop type: Both-IRC+RFC
Below normal (%)
40
55
70
85
100
Normal (%)
40
55
70
85
100
Above normal (%)
40
55
70
85
100

Figure 6.5: Probability of the most likely tercile for precipitation multi-model predictions for crop seasons of 2023-2026 (start date 2021, crop seasons 2-5) over the global cotton crop regions. The intensity of colours indicate the probability for each tercile category. White colours show that all categories are equiprobable (all of them with a probability less than 40%).

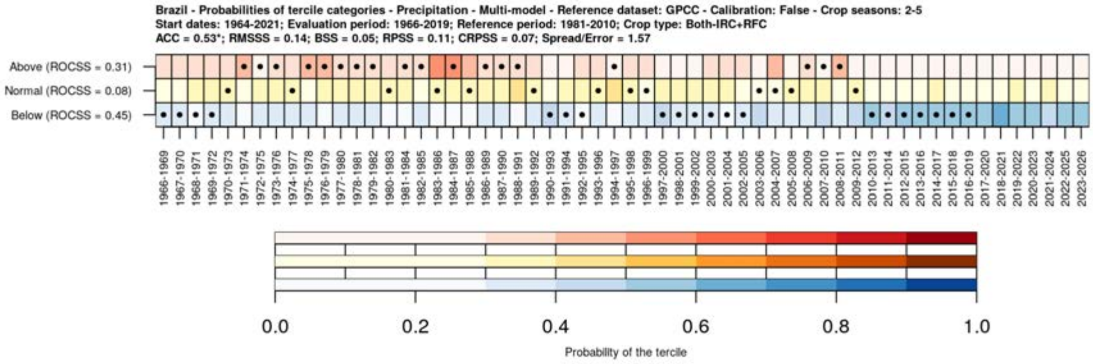


Figure 6.6: Probability of tercile categories for precipitation multi-model predictions for crop seasons 2-5 over Brazil. The intensity of colours indicate the probability for each tercile category. Dots show the observed category. The skill estimates are shown in the title.

Chapter 7

Conclusions

The research conducted during this Ph.D. thesis aimed at evaluating the current generation of decadal climate predictions, identifying windows of opportunity for which decadal forecasts show skill in predicting climate variations so that they can potentially be used for climate services to support decision-making and adaptation strategies in socio-economic sectors that are affected by climate variability and change.

The work has focused on evaluating the CMIP6 decadal predictions quality (measured as the degree of agreement between predictions and observations of the past climate conditions since 1961) to estimate how decadal forecast systems are expected to predict future variations of the climate system. However, it should be noted that hindcast skill does not imply forecast skill since the drives and sources of predictability may change. Besides, several post-processing techniques have been applied to the raw predictions to enhance the forecast quality, reliability and usability. These post-processing techniques include the multi-model combination, correction of systematic model biases by calibration methods, and computation of weather regimes and extreme climate indicators. In addition, the skill of decadal predictions has been compared to that of historical forcing simulations to estimate the impact of model initialisation towards the observed climate state. All these comparisons allow for selecting the best source of climate information for next year up to one decade.

This chapter presents a summary of the main conclusions (Section 7.1) and recommendations for future research in the field of decadal prediction (Section 7.2).

7.1. Main conclusions

- The CMIP6 decadal predictions are skilful in predicting changes in mean near-surface temperature over most of the globe at multi-annual time scales. Extreme climate events associated with daily maximum and minimum temperature are also well predicted at multi-annual time scales, particularly those based on minimum temperature. Also, extremes based on frequency are generally predicted with higher skill than those based on intensity. On the other hand, predictions of mean and extreme precipitation show skill over limited regions, thus limiting their potential usability for climate services provision. AMV index and the GSAT anomalies predictions have also been found to be skilful. Nevertheless, the multi-model ensemble overestimates the increase in GSAT anomalies in recent years. Calibration methods correct this overestimation but with the drawback of reduced skill in the 1960s and 1970s.

- The forecast quality, as measured with bias-sensitive metrics (such as the RMSSS and CRPSS), is enhanced when applying bias-adjustment and calibration techniques, as systematic biases are partially corrected through calibration. Other skill metrics that are not sensitive to biases in mean and variance (such as ACC and RPSS) are generally decreased due to the application of these methods in cross-validation mode. However, cross-validation is essential when bias-adjusting or calibrating hindcasts not to produce overfitting and thus overestimate the actual skill of future predictions.

- The multi-model ensemble is the most reasonable approach when systematically providing predictions (e.g. globally for several variables), even if the multi-model ensemble does not outperform the best individual forecast system for each case. On the other hand, the most skilful model or set of models within the multi-model ensemble could be selected to deliver the best climate information for each specific case under consideration (e.g. a particular variable over a limited region).

- The ensemble size of the multi-model ensemble impacts the forecast quality, as found when comparing a research multi-model ensemble (built with predictions produced with 13 forecast systems which provide hindcasts) against an operational multi-model (built with predictions from 4 forecast systems providing forecasts in addition to hindcasts). This implies a major barrier to the development of climate services, as the limited number of forecast systems providing near-real-time forecasts decreases the quality (and thus usability) of predictions compared to the quality estimated using all decadal prediction systems. Besides, more real-time predictions allow the selection of the best forecast system or multi-model sub-ensemble for each specific region, variable and forecast period.

- The EC-Earth3 forecast system correctly represents the spatial patterns and climatological occurrence frequencies of the four Euro-Atlantic weather regimes. However, the decadal forecast system does not skillfully predict the inter-annual to decadal variations of the occurrence frequencies, and there are no significant skill differences compared to the uninitialised historical forcing simulations.

- There is an added value from model initialisation for multi-model temperature and precipitation predictions over some ocean and land regions, in addition to the AMV index and GSAT anomalies. Such added value is generally low and highly region-dependent for temperature and precipitation extremes predictions.

- More co-production and sharing of knowledge is needed from both user and scientist sides in order to produce more tailored and usable climate information to suitably underpin decision-making processes.

7.2. Future perspectives

The work developed during this Ph.D. thesis serves as the basis for future research and application of decadal predictions for climate services in different socio-economic sectors impacted by climate variability and change.

Apart from the variables considered in the research articles and contributions to projects within this thesis, other variables and user-oriented indicators should be explored to assess their predictability and potential use for particular applications. To do so, it is necessary to interact more with final users, as the co-development process is essential to produce a final product tailored and usable by stakeholders in a decision-making process. Also, having a more comprehensive range of users will allow for gaining more knowledge and transferring it to future climate services provision.

Based on the results of the studies developed within this Ph.D. thesis, future studies could address the capacity of decadal models' output to be used as inputs for statistical models, which may be skilful in predicting more user-relevant variables such as crop yield for the agriculture sector or the capacity factor for the renewable energy sector, instead of essential climate variables that may be of less potential usage for users.

These statistical models can also be built with large-scale indicators such as the NAO index. However, one requirement is that the large-scale indicators should be skillfully predicted. Smith *et al.* (2020) recently showed that NAO predictions suffer from the signal-to-noise issue, and that this limitation could be addressed by exploiting very large ensembles and sub-selecting ensemble members with the correct NAO magnitude. This methodology could be further developed to get improved skill, and could also be applied to other variables and large-scale indicators.

The coarse resolution and biases of climate models are also a barrier to utilising the predictions for local decisions. There are a number of calibration and downscaling approaches that aim to correct such biases and regionalise the predictions. However, the development of more complex statistical techniques might outperform current methods. If they show effectiveness when applied to decadal forecasts, these new methodologies will increase the reliability and usability of climate information to support regional applications.

The systematic forecast quality assessment performed in this Ph.D. thesis also serves as a benchmark to compare the skill of current and future forecast systems and to provide climate modellers with weak points of the current models so that they can put efforts into improving such aspects. Future research is also needed to understand the sources of predictability of skill at annual to decadal time scales, which will enable the improvement of current climate models, thus enhancing the forecast quality and reliability. Fixing essential issues such as model biases and the low signal-to-noise ratio present in current model ensembles will enable the utilisation of climate information extracted from climate models for decision-making at annual to decadal time scales.

Another question that should be addressed is how much skill is sufficient for users to base their decision on climate predictions. Different users might have different thresholds to consider a prediction as "skilful enough" depending on their particular decisions. Also, even if the predictions do not show skill over the entire evaluation period, some climate variations might be more predictable depending on the specific state of the drivers or sources of predictability. For instance, the skill could differ depending on the phase of some modes of variability. In such a case, more advanced knowledge of those drivers would enhance the trustworthiness of the forecasts issued during a particular climate state.

Appendix A

Contributions and dissemination

A.1. Authored papers

- **Delgado-Torres, C.**, Donat, M. G., Gonzalez-Reviriego, N., Caron, L., Athanasiadis, P. J., Bretonnière, P., Dunstone, N. J., Ho, A., Nicolì, D., Pankatz, K., Paxian, A., Pérez-Zanón, N., Cabré, M. S., Solaraju-Murali, B., Soret, A., and Doblas-Reyes, F. J. (2022). Multi-Model forecast quality assessment of CMIP6 decadal predictions. Journal of Climate, 35(13), 4363-4382. https://doi.org/10.1175/JCLI-D-21-0811.1

- **Delgado-Torres, C.**, Verfaillie, D., Mohino, E., and Donat, M. G. (2022). Representation and annual to decadal predictability of Euro-Atlantic weather regimes in the CMIP6 version of the EC-Earth coupled climate model. Journal of Geophysical Research: Atmospheres, 127, e2022JD036673. https://doi.org/10.1029/2022JD036673

- **Delgado-Torres, C.**, Donat, M. G., Soret, A., Gonzalez-Reviriego, N., Bretonnière, P.-A., Ho, A.-C., Pérez-Zanón, N., Samsó Cabré, M., and Doblas-Reyes, F. J. (2023). Multi-annual predictions of the frequency and intensity of daily temperature and precipitation extremes. Environmental Research Letters, 18 034031. https://doi.org/10.1088/1748-9326/acbbe1

A.2. Co-authored papers

- De Luca, P., **Delgado-Torres, C.**, Mahmood, R., Samso-Cabre, M., and Donat, M. G. (2023). Constraining decadal variability regionally improves near-term projections of hot, cold and dry extremes. Environmental Research Letters, 18 094054. https://doi.org/10.1088/1748-9326/acf389

- De Luca, P., **Delgado-Torres, C.**, Mahmood, R., Samso-Cabre, M., Keenlyside, N. S., and Donat, M. G. (in prep). Decadal predictability of summer hot, dry and compound hot-dry extremes in southeastern Australia, Environmental Research: Climate.

- Donat, M. G., **Delgado-Torres, C.**, De Luca, P., Ortega, P., and Doblas-Reyes, F. J. (2023). How Credibly Do CMIP6 simulations capture historical mean and extreme precipitation changes? Geophysical Research Letters, 50, e2022GL102466. https://doi.org/10.1029/2022GL102466

- Liu, Y., Donat, M. G., England, M. H., Alexander, L. V., Hirsch, A. L., and **Delgado-Torres, C.** (2023). Enhanced multi-year predictability after El Niño and La Niña events. Nature Communications, 14, 6387. https://doi.org/10.1038/s41467-023-42113-9

- Martínez-Boti, A., Solaraju-Murali, B., Marcos, R., Manrique-Suñén, A., Gonzalez-Reviriego, N., **Delgado-Torres, C.**, and Soret, A. (in prep). Exploring windows of opportunity for the use of seasonal predictions of droughts.

- Milders, N., Agudetse, V., Bretonnière, P.-A., **Delgado-Torres, C.**, Gonzalez-Reviriego, N., Pérez-Zanón, N., Rifà, E., Samsó Cabré, M., and Doblas-Reyes, F. J. (in prep). Comprehensive assessment of multi-model seasonal climate forecasts.

- Moreno-Montes, S., **Delgado-Torres, C.**, Marcos, R., Ramon, J., Duzenli, E., and Soret, A. (in prep). A comparative analysis of multi-model and downscaled decadal climate predictions over the Southern African Development Community.

- Pérez-Zanón, N., Caron, L.-P., Terzago, S., Van Schaeybroeck, B., Lledó, L., Manubens, N., Roulin, E., Alvarez-Castro, M. C., Batté, L., Bretonnière, P.-A., Corti, S., **Delgado-Torres, C.**, Domínguez, M., Fabiano, F., Giuntoli, I., von Hardenberg, J., Sánchez-García, E., Torralba, V., and Verfaillie, D. (2022). Climate Services Toolbox (CSTools) v4.0: from climate forecasts to climate forecast information. Geoscientific Model Development, 15, 6115–6142. https://doi.org/10.5194/gmd-15-6115-2022

- Pérez-Zanón, N., Agudetse, V., Baulenas, E., Bretoniere, P.-A., **Delgado-Torres, C.**, González-Reviriego, N., Manrique-Suñén, A., Nicodemou, A., Olid, M., Palma, Ll., Terrado, M., Basile, B., Carteni, F., Dente, A., Esquerra, C., Olani, F., Otero, M., Santos-Alves, F., Torres, M., Valente, J., and Soret, A. (in prep). Lessons learned from the co-development of operational climate forecast services for vineyards management. Climate Risk Management Journal.

- Mahmood, R., Donat, M. G., Ortega, P., Doblas-Reyes, F. J., **Delgado-Torres, C.**, Samsó, M., and Bretonnière, P.-A. (2022). Constraining low-frequency variability in climate projections to predict climate on decadal to multi-decadal timescales – a poor man's initialized prediction system. Earth System Dynamics, 13, 1437–1450, https://doi.org/10.5194/esd-13-1437-2022

- Solaraju-Murali, B., Doblas-Reyes, F. J., Torralba, V., **Delgado-Torres, C.**, and González-Reviriego, N. (in prep). Assessing seamless seasonal to decadal predictions for supporting climate change adaptation.

A.3. Conferences

- 9th BSC Doctoral Symposium 2022 (9BSCDS): **Delgado-Torres, C.**, Donat, M. G., and Soret, A.: Multi-model Forecast Quality Assessment of CMIP6 Decadal Predictions. 10th-12th May, 2022. Barcelona, Spain. https://www.bsc.es/education/predoctoral-phd/doctoral-symposium/9th-bsc-doctoral-symposium-2022/agenda

- 9th BSC Doctoral Symposium 2022 (9BSCDS): Ramon, J., Lledó, Ll., Palma, Ll., **Delgado-Torres, C.**, and Marcos, R.: CSDownscale: an R Package for Statistical Downscaling. 10th-12th May, 2022. Barcelona, Spain. https://www.bsc.es/education/predoctoral-phd/doctoral-symposium/9th-bsc-doctoral-symposium-2022/agenda

- EGU General Assembly 2022 (EGU22): **Delgado-Torres, C.**, Donat, M. G., Gonzalez-Reviriego, N., Caron, L.-P., Athanasiadis, P. J., Bretonnière, P.-A., Dunstone, N. J., Ho, A.-C., Pankatz, K., Paxian, A., Pérez-Zanón, N., Samsó Cabré, M., Solaraju-Murali, B., Soret, A., and Doblas-Reyes, F. J.: Multi-model forecast quality assessment of CMIP6 decadal predictions. 23th-27th May, 2022. Vienna, Austria. https://doi.org/10.5194/egusphere-egu22-13156

- 3rd WMO Workshop on Operational Climate Prediction (WMO OCP-3): **Delgado-Torres, C.**, Donat, M. G., Gonzalez-Reviriego, N., Caron, L.-P., Athanasiadis, P. J., Bretonnière, P.-A., Dunstone, N. J., Ho, A.-C., Pankatz, K., Paxian, A., Pérez-Zanón, N., Samsó Cabré, M., Solaraju-Murali, B., Soret, A., and Doblas-Reyes, F. J.: Multi-model forecast quality assessment of CMIP6 decadal predictions. 20th-22nd September, 2022. Lisbon, Portugal. https://community.wmo.int/meetings/ocp-3

- AGU Fall Meeting 2022 (AGU22): **Delgado-Torres, C.**, Donat, M. G., Soret, A., Gonzalez-Reviriego, N., Bretonnière, P., Ho, A., Pérez-Zanón, N., Cabré, M. S., and Doblas-Reyes, F. J.: Decadal Prediction Skill for Daily Temperature and Precipitation Extreme Climate Events. 12th-16th December, 2022. Chicago, USA. https://agu.confex.com/agu/fm22/meetingapp.cgi/Paper/1055746

- EGU General Assembly 2023 (EGU23): **Delgado-Torres, C.**, Donat, M. G., Soret, A., González-Reviriego, N., Bretonnière, P.-A., Ho, A.-C., Pérez-Zanón, N., Samsó Cabré, M., and Doblas-Reyes, F. J.: Multi-annual predictions of daily temperature and precipitation extremes: forecast quality and impact of model initialisation. 24th-28th April, 2023. Vienna, Austria. https://doi.org/10.5194/egusphere-egu23-2399

- EGU General Assembly 2023 (EGU23): Doblas-Reyes, F. J., Agudetse, V., **Delgado-Torres, C.**, Donat, M. G., González-Reviriego, N., De Luca, P., Milders, N., G. Muñoz, A., Palma, L., Pérez-Zanón, N., Ramon, J., Solaraju-Murali, B., Soret, A., and Torralba, V.: Forecast quality of climate extreme predictions and its relevance for climate services. 24th-28th April, 2023. Vienna, Austria. https://doi.org/10.5194/egusphere-egu23-11143

- 10th BSC Doctoral Symposium 2023 (10BSCDS): **Delgado-Torres, C.**, Donat, M. G., and Soret, A.: Multi-annual predictions of daily temperature and precipitation extremes. 9th-10th May, 2023. Barcelona, Spain. https://www.bsc.es/education/predoctoral-phd/doctoral-symposium/10th-international-bsc-severo-ochoa-doctoral-symposium-2023

- XXVIII General Assembly of the International Union of Geodesy and Geophysics (IUGG 2023): Donat, M., De Luca, P., **Delgado-Torres, C.**, Mahmood, R.: Multi-decadal predictability of wet and dry precipitation extremes from external forcing and climate variability. 11th-20th July, 2023. Berlin, Germany. https://doi.org/10.57757/IUGG23-3485

- World Climate Research Programme Open Science Conference 2023 (WCRP OSC 2023): Bojovic, D., Octenjak, S., **Delgado-Torres, C.**, Vigo, I., Marcos, R.: From local knowledge to climate science: co-creating climate information across time and spatial scales for food security in Malawi. 23rd-27th October, 2023. Kigali, Rwanda. https://wcrp-osc2023.org/

- World Climate Research Programme Open Science Conference 2023 (WCRP OSC 2023): Donat, M. G., Mahmood, R., De Luca, P., **Delgado-Torres, C.**, Cos, J., Ortega, P., Doblas-Reyes, F. J.: Constraining decadal variability in large climate projections ensembles to obtain improved near-term climate change estimates and attribute sources of predictability. 23rd-27th October, 2023. Kigali, Rwanda. https://wcrp-osc2023.org/

- World Climate Research Programme Open Science Conference 2023 (WCRP OSC 2023): Donat, M. G., De Luca, P., **Delgado-Torres, C.**, Mahmood, R.: Multi-decadal predictability of wet and dry precipitation extremes from external forcing and climate variability. 23rd-27th October, 2023. Kigali, Rwanda. https://wcrp-osc2023.org/

A.4. Workshops and hackathons

- C3S_34c contract: Workshop on decadal predictions data standards. 8th and 9th Jun 2020.

- EUCP project: Workshop on FAIR data and software. 28th Oct, 4th Nov and 11th Nov 2020. http://doi.org/10.5281/zenodo.4279433
- IS-ENES3 Workshop on Climate Indices: Eastern Europe perspective: 17th May 2021.
- WCRP: Workshop on Extremes in Climate Prediction Ensembles (ExCPEns): 25th-27th Oct 2021. https://trello.com/b/0h9RHCb4/wcrp-excpens-workshop-2021
- EUCP project: EUCP Final Multi-User Forum workshop: 3rd May 2022. https://www.eucp-project.eu/eucp-updates/eucp-final-multi-user-forum-online-event-3rd-may-2022/
- EUCP project: EUCP Final meeting, 4th-6th May 2022. https://www.eucp-project.eu/eucp-updates/eucp-final-meeting/
- NextGEMS project: Hackathon for the renewable energy sector: June 28th - July 2nd 2022. https://indico.mpimet.mpg.de/event/41/overview
- Columbia University: Storytelling 101 with Terri Trespicio: October 12th 2022. https://events.columbia.edu/cal/event/showEventMore.rdo;jsessionid=ekXzyCaf4b3KRIo1sireauzt-PNfts-l991-qTzA.calprdapp05
- Columbia University: Soccer in a Warming World Workshop: November 16th 2022. https://www.eventbrite.com/e/soccer-in-a-warming-world-workshop-tickets-443399769647
- NextGEMS project: Hackathon for the fisheries sector: May 29th - June 2nd 2023. https://events.mpimet.mpg.de/event/56/

A.5. Ph.D. research stay

A research visit was conducted at the International Research Institute for Climate and Society (IRI; https://iri.columbia.edu/) at Columbia University in the City of New York, NY, USA, from September 16th 2022 to December 15th 2022.

During the secondment, a research work was carried out on the evaluation of the predictability of extreme climate events at inter-annual to decadal timescales, and their potential applications for climate services, under the supervision of Dr. Ángel G. Muñoz and Carmen González Romero.

This work was part of the research article published in Environmental Research Letters (Chapter 5), and served as the basis of the following climate services oriented research. In addition, it was a great opportunity to know and interact with excellent scientists of the field, as well as to learn how another key climate research institution works.

A.6. Contribution to software development

- CSDownscale (not yet on CRAN): R-based software package intended for downscaling climate predictions. So far, only purely statistical methods are included. The downscaling can be performed either to a grid of different spatial resolution or to a point location.

- CSScorecards (not yet on CRAN): R-based software package to create scorecards, which are useful tools for visualisation of systematic climate forecast verification metrics.

- CSTools (https://CRAN.R-project.org/package=CSTools): The Climate Services Tools, CSTools, is an R package designed and built to assess and improve the quality of climate forecasts for seasonal to multi–annual scales. The package contains process-based state-of-the-art methods for forecast calibration, bias correction, statistical and stochastic downscaling, optimal forecast combination and multivariate verification, as well as basic and advanced tools to obtain tailored products.

- s2dv (https://CRAN.R-project.org/package=s2dv): R-based software package intended for seasonal-to-decadal climate forecast verification. This package is specially designed for the comparison between the experimental and observational datasets. The functionality of the included functions covers data retrieval, data post-processing, skill scores against observation, and visualisation.

- SUNSET (SUbseasoNal to decadal climate forecast post-processIng and asSEssmenT suite): R-based tool that provides climate services for sub-seasonal, seasonal and decadal climate forecast horizons. The tool post-processes climate forecast outputs by applying state-of-the-art methodologies to tailor climate products for each application and sector (e.g.: agriculture, energy, water management, or health). Its modular design allows the technicians and researchers to decide on the post-processing required steps, such as regridding, anomalies, downscaling, bias-adjustment methods, as well as the products definition by deciding on the forecast system and reference datasets, variables, and forecast horizon among others. The tool also allows the creation and visualisation of climate forecast products, such as maps for the most likely terciles, and performs the verification of the products using user-defined metrics, which can be visualised on maps and scorecards. The integration of Autosubmit (Python-based workflow manager to create, manage and monitor complex tasks involving different substeps) in the tool allows users to parallelize the computation in High Performance Computing (HPC) machines.

A.7. Contribution to projects

The outcomes and software developed within this PhD thesis have been applied to several Spanish and European projects, as well as to a contract with a private company.

- CLINSA project
  - Predicción decadal climática para servicios climáticos a corto plazo y adaptación (https://www.bsc.es/research-and-development/projects/clinsa-prediccion-decadal-climatica-para-servicios-climaticos)
  - Funding: Ministerio de Ciencia, Innovación y Universidades (CGL2017-85791-R)
  - Task: Illustration of the relative merits of the calibration, combination and downscaling of the decadal predictions

- C3S_34c contract
  - Prototype Service for Decadal Climate Predictions (https://climate.copernicus.eu/c3s34c-prototype-service-decadal-climate-predictions)
  - Funding: Copernicus Climate Change Service (C3S) operated by the European Centre for Medium-Range Weather Forecasts (ECMWF) contract ECMWF/COPERNICUS/2019/C3S_34c_DWD
  - Task: Recommendations for forecast quality assessment and forecast product generation
- EUCP project
  - European Climate Prediction system (https://www.eucp-project.eu/)
  - Funding: European Union under Horizon 2020 Programme under grant agreement 776613
  - Tasks: Evaluation of the representation and predictability of Euro-Atlantic weather regimes; Forecast quality assessment of the CMIP6 DCPP multi-model ensemble; Case study on the wind energy sector
- Decathlon contract
  - Decathlon (https://www.decathlon.es/)
  - Task: Forecast quality assessment and forecast provision of climate variables related to cotton production with focus on the specific regions and seasons
- FOCUS-Africa project
  - Full-value chain Optimised Climate User-centric Services for Southern Africa (https://focus-africaproject.eu/)
  - Funding: European Union under Horizon 2020 Programme under grant agreement 869575

  - Tasks: Evaluation of individual models over the SADC; Comparison of different calibration techniques; Comparison of different multi-model approaches; Comparison of different downscaling approaches; Evaluation of SPEI predictions; Case study on the agriculture sector using downscaled decadal predictions of climate variables and indices related to maize production

- ASPECT project

  - ASPECT Facilitating Seamless Climate Adaptation (https://www.aspect-project.eu/)
  - Funding: European Union under Horizon Europe grant agreement 101081460
  - Task: Forecast quality assessment of extreme indices based on daily temperature and precipitation

Appendix B

Supplementary material for Chapter 3

Delgado-Torres, C., Donat, M. G., Gonzalez-Reviriego, N., Caron, L., Athanasiadis, P. J., Bretonnière, P., Dunstone, N. J., Ho, A., Nicoli, D., Pankatz, K., Paxian, A., Pérez-Zanón, N., Cabré, M. S., Solaraju-Murali, B., Soret, A., and Doblas-Reyes, F. J. (2022). Multi-Model Forecast Quality Assessment of CMIP6 Decadal Predictions. Journal of Climate, 35(13), 4363-4382. https://doi.org/10.1175/JCLI-D-21-0811.1

Main objectives, main outcomes and research article in Chapter 3.

Supplemental Material for Multi-model forecast quality assessment of CMIP6 decadal predictions

Carlos Delgado-Torres,[a] Markus G. Donat,[a,b] Nube Gonzalez-Reviriego,[a] Louis-Philippe Caron,[a,c] Panos J. Athanasiadis,[d] Pierre-Antoine Bretonnière,[a] Nick J. Dunstone,[e] An-Chi Ho,[a] Dario Nicolì, [d] Klaus Pankatz,[f] Andreas Paxian,[f] Núria Pérez-Zanón,[a] Margarida Samsó Cabré,[a] Balakrishnan Solaraju-Murali,[a] Albert Soret,[a] and Francisco J. Doblas-Reyes[a,b]

[a] *Barcelona Supercomputing Center (BSC), Barcelona, Spain*
[b] *Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain*
[c] *Ouranos, 550 Sherbrooke St W, Montreal, QC, Canada*
[d] *Centro Euro-Mediterraneo sui Cambiamenti Climatici, Bologna, Italy*
[e] *Met Office Hadley Centre, FitzRoy Road, Exeter EX1 3PB, UK*
[f] *Business Area of Climate and Environment, Deutscher Wetterdienst, Offenbach (Main), Germany*

ABSTRACT: Decadal climate predictions are a relatively new source of climate information for inter-annual to decadal time scales, which is of increasing interest for users. Forecast quality assessment is essential to identify windows of opportunity (e.g., variables, regions, and forecast periods) with skill that can be used to develop climate services to inform users in several sectors and define benchmarks for improvements in forecast systems. This work evaluates the quality of multi-model forecasts of near-surface air temperature, precipitation, Atlantic multi-decadal variability index (AMV) and global near-surface air temperature anomalies (GSAT) generated from all the available retrospective decadal predictions contributing to the Coupled Model Intercomparison Project Phase 6 (CMIP6). The predictions generally show high skill in predicting temperature, AMV, and GSAT, while the skill is more limited for precipitation. Different approaches for generating a multi-model forecast are compared, finding small differences between them. The multi-model ensemble is also compared to the individual forecast systems. The best system usually provides the highest skill. However, the multi-model ensemble is a reasonable choice for not having to select the best system for each particular variable, forecast period and region. Furthermore, the decadal predictions are compared to the historical simulations to estimate the impact of initialization. An added value is found for several ocean and land regions for temperature, AMV, and GSAT, while it is more reduced for precipitation. Moreover, the full ensemble is compared to a sub-ensemble to measure the impact of the ensemble size. Finally, the implications of these results in a climate services context, which requires predictions issued in near real-time, are discussed.

Corresponding author: Carlos Delgado-Torres, carlos.delgado@bsc.es

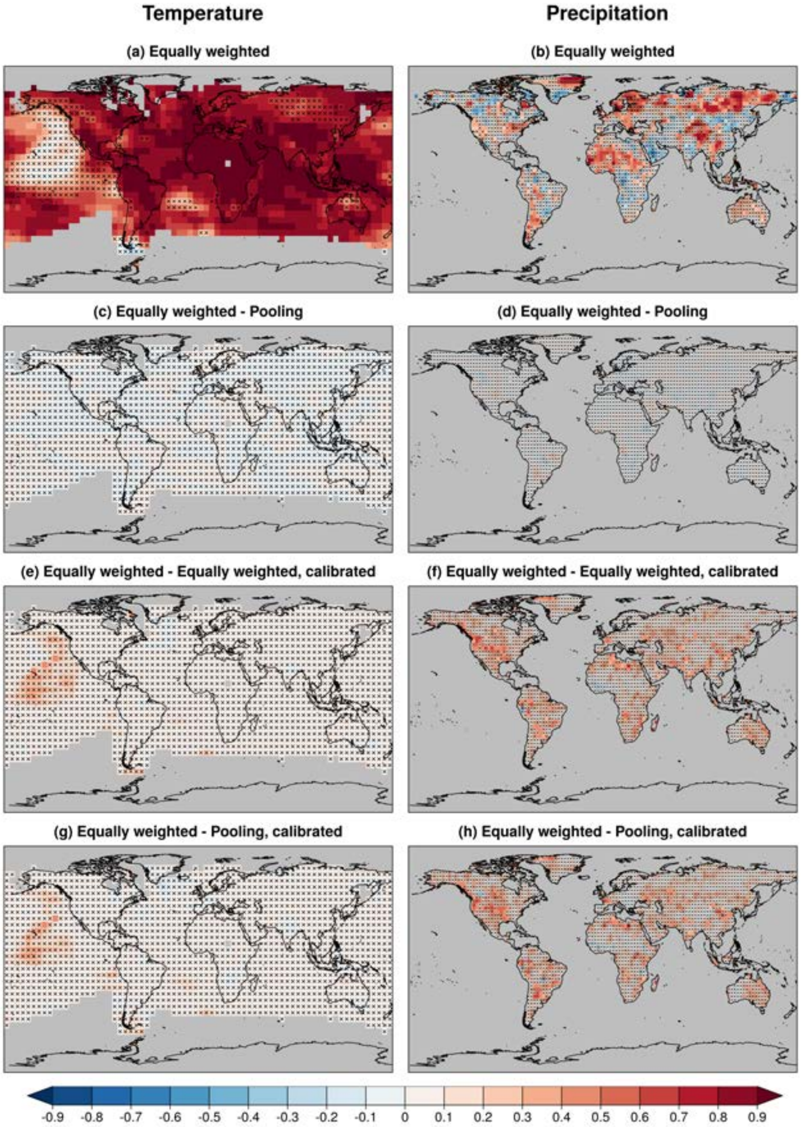

Temperature
Precipitation
(a) Equally weighted
(b) Equally weighted
(c) Equally weighted - Pooling
(d) Equally weighted - Pooling
(e) Equally weighted - Equally weighted, calibrated
(f) Equally weighted - Equally weighted, calibrated
(g) Equally weighted - Pooling, calibrated
(h) Equally weighted - Pooling, calibrated
-0.9
-0.8
-0.7
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

FIG. S1. Maps of ACC obtained with the multi-model-1 approach (first row) and maps of ACC differences between the multi-model-1 and the rest of the multi-model approaches (second to fourth rows) for the forecast years 1–5 for the surface air temperature (first column) and precipitation (second column). The ACC has been computed over the 1961–2014 period (start dates 1960–2009) for each individual grid point. The reference period for the computation of anomalies is 1981–2010. The reference datasets used for the surface air temperature and precipitation are the GHCNv4 and the GPCC datasets, respectively. Crosses indicate that the values are not statistically significant at the 95% level using a two-sided t-test accounting for autocorrelation.

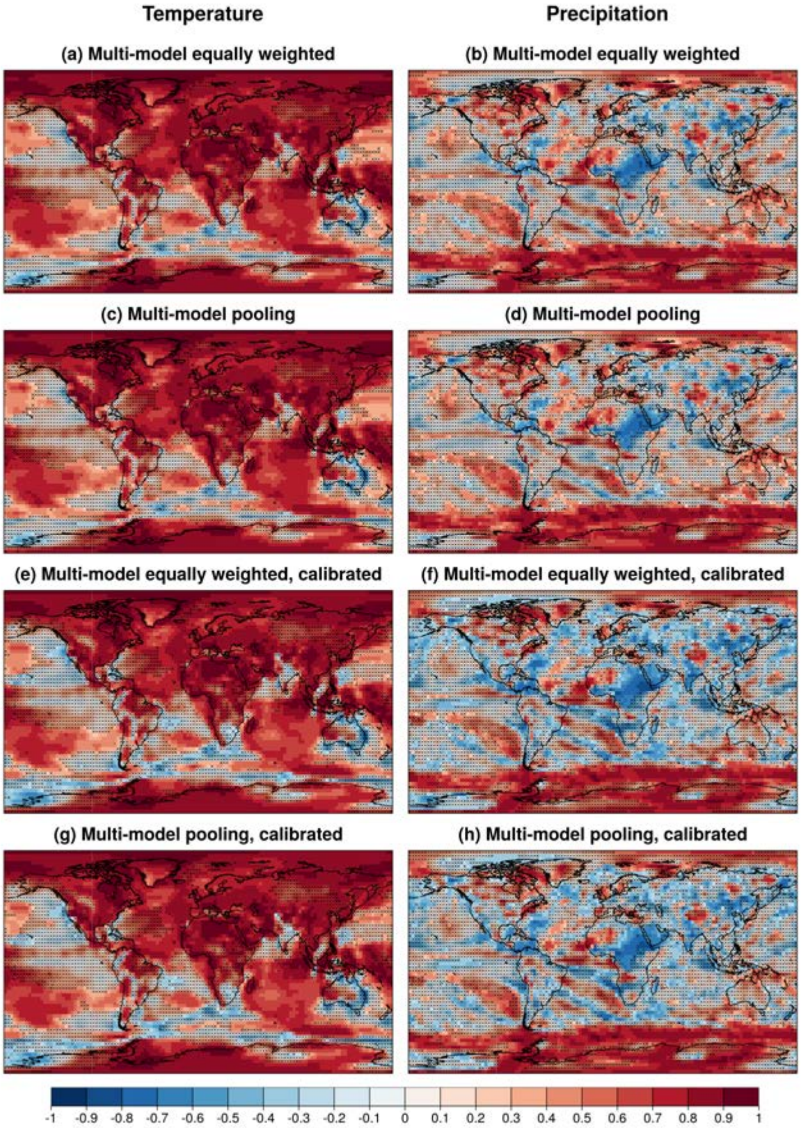

Temperature
Precipitation
(a) Multi-model equally weighted
(b) Multi-model equally weighted
(c) Multi-model pooling
(d) Multi-model pooling
(e) Multi-model equally weighted, calibrated
(f) Multi-model equally weighted, calibrated
(g) Multi-model pooling, calibrated
(h) Multi-model pooling, calibrated
-1
-0.9
-0.8
-0.7
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

FIG. S2. Same as Figure 1, but using the JRA-55 reanalysis as the reference dataset.

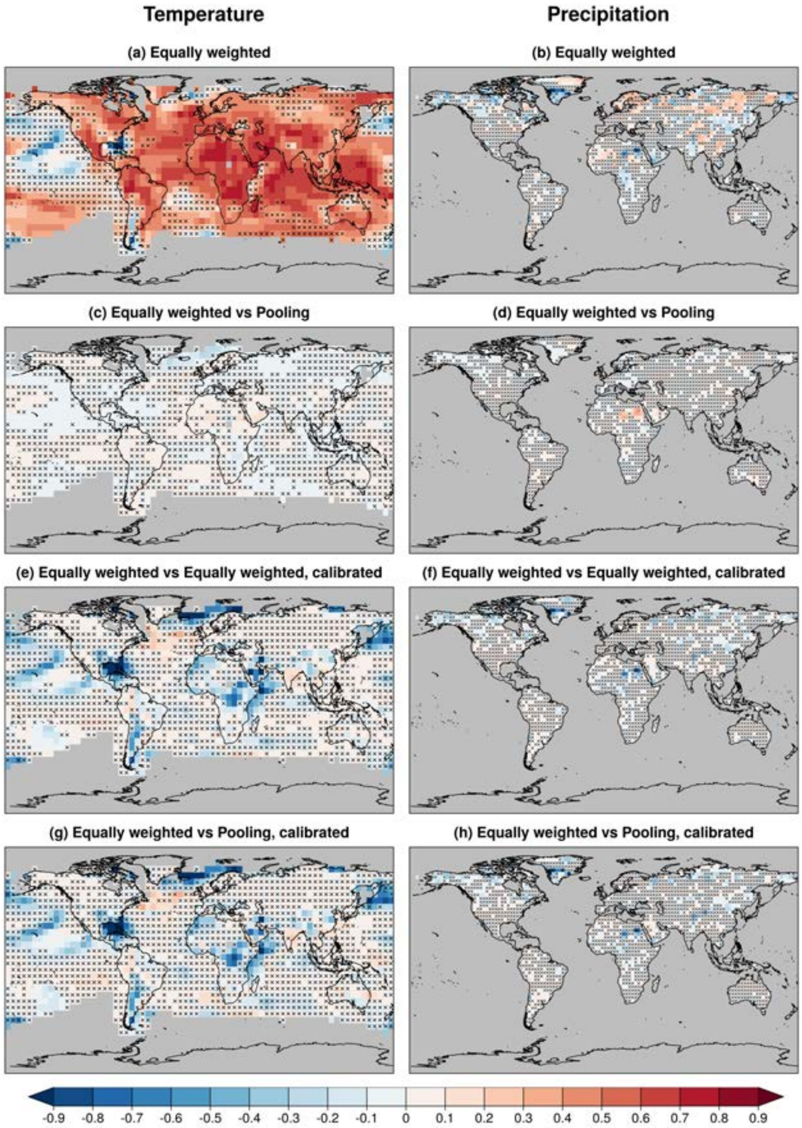

Temperature
Precipitation
(a) Equally weighted
(b) Equally weighted
(c) Equally weighted vs Pooling
(d) Equally weighted vs Pooling
(e) Equally weighted vs Equally weighted, calibrated
(f) Equally weighted vs Equally weighted, calibrated
(g) Equally weighted vs Pooling, calibrated
(h) Equally weighted vs Pooling, calibrated
-0.9 -0.8 -0.7 -0.6 -0.5 -0.4 -0.3 -0.2 -0.1 0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9

FIG. S3. Maps of RMSSS obtained with the multi-model-1 approach using the climatology as the reference forecast (first row) and maps of RMSSS of the multi-model-1 using the rest of multi-model approaches as the reference forecast (second to fourth rows) for the forecast years 1–5 for near-surface air temperature (first column) and precipitation (second column). The skill scores have been computed over the 1961–2014 period (start dates 1960–2009). The reference period for the computation of the anomalies is 1981–2010. The reference datasets used for the surface air temperature and precipitation are the GHCNv4 and the GPCC datasets, respectively. Crosses indicate that the decadal predictions do not provide significantly better or worse predictions than the reference forecast at the 95% confidence level based on a Random Walk test.

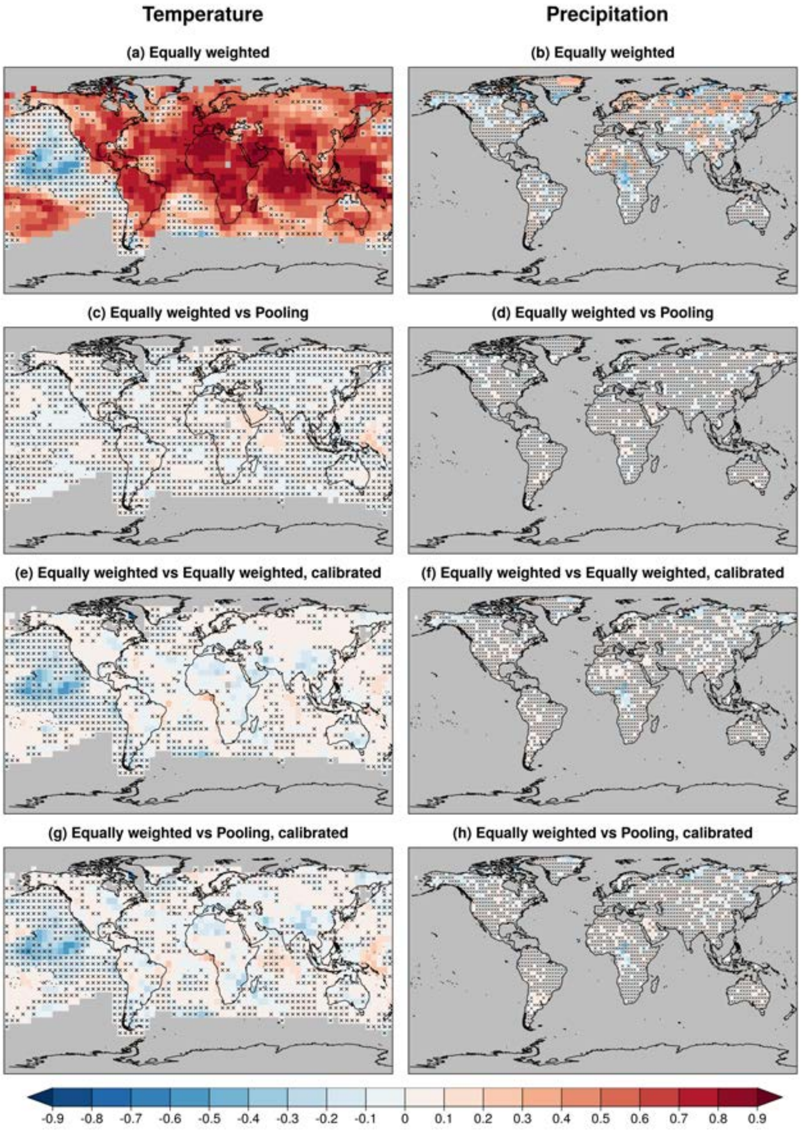

Temperature
Precipitation
(a) Equally weighted
(b) Equally weighted
(c) Equally weighted vs Pooling
(d) Equally weighted vs Pooling
(e) Equally weighted vs Equally weighted, calibrated
(f) Equally weighted vs Equally weighted, calibrated
(g) Equally weighted vs Pooling, calibrated
(h) Equally weighted vs Pooling, calibrated
-0.9 -0.8 -0.7 -0.6 -0.5 -0.4 -0.3 -0.2 -0.1 0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9

Fig. S4. Maps of RPSS for 3 categories obtained with the multi-model-1 approach using the climatology as the reference forecast (first row) and maps of RPSS for 3 categories of the multi-model-1 using the rest of multi-model approaches as the reference forecast (second to fourth rows) for the forecast years 1–5 for near-surface air temperature (first column) and precipitation (second column). The skill scores have been computed over the 1961–2014 period (start dates 1960–2009). The reference period for the computation of the thresholds between categories is 1981–2010. The reference datasets used for the surface air temperature and precipitation are the GHCNv4 and the GPCC datasets, respectively. Crosses indicate that the decadal predictions do not provide significantly better or worse predictions than the reference forecast at the 95% confidence level based on a Random Walk test.

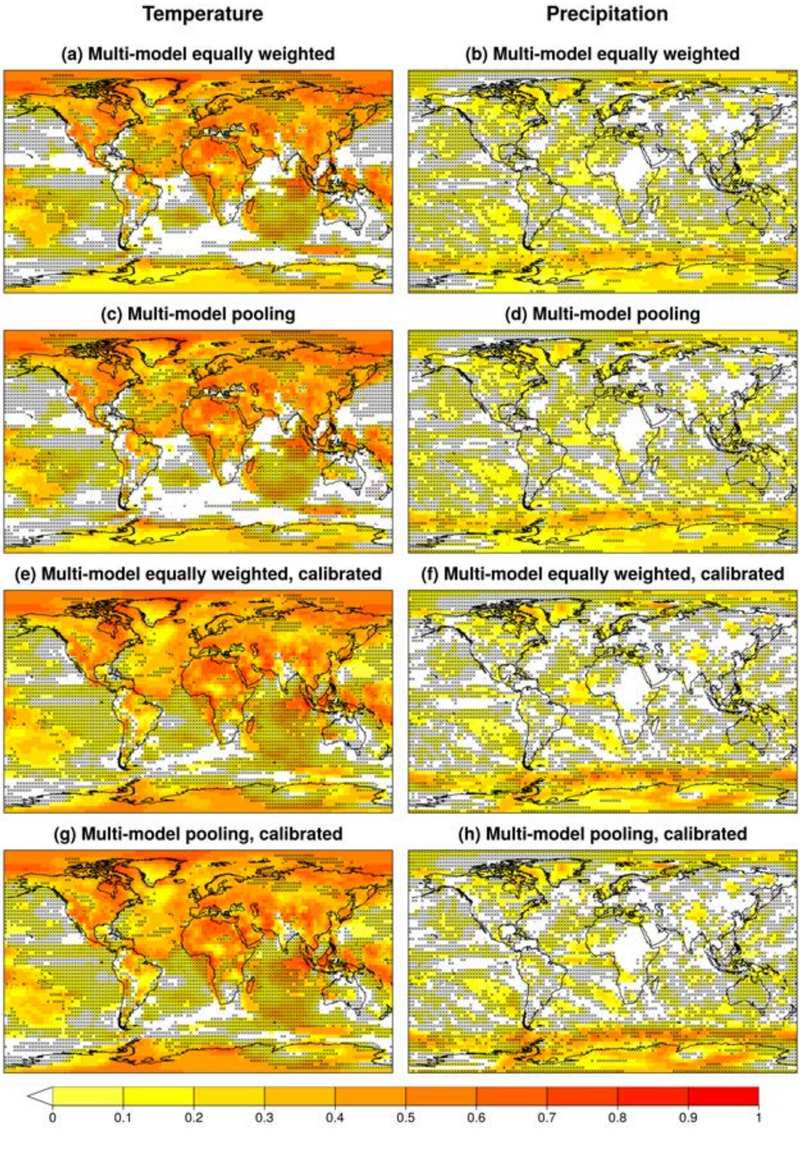

Temperature
Precipitation
(a) Multi-model equally weighted
(b) Multi-model equally weighted
(c) Multi-model pooling
(d) Multi-model pooling
(e) Multi-model equally weighted, calibrated
(f) Multi-model equally weighted, calibrated
(g) Multi-model pooling, calibrated
(h) Multi-model pooling, calibrated
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

FIG. S5. Same as Figure 2, but using the JRA-55 reanalysis as the reference dataset.

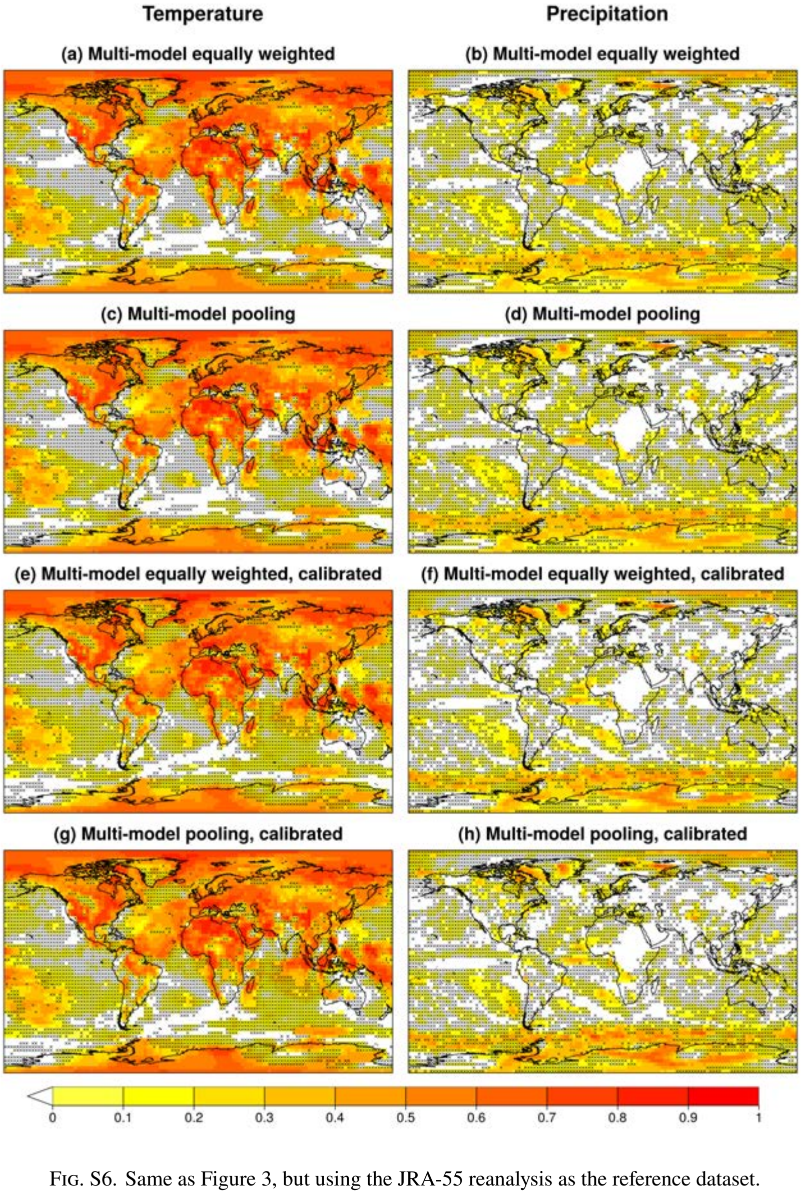

Temperature
Precipitation
(a) Multi-model equally weighted
(b) Multi-model equally weighted
(c) Multi-model pooling
(d) Multi-model pooling
(e) Multi-model equally weighted, calibrated
(f) Multi-model equally weighted, calibrated
(g) Multi-model pooling, calibrated
(h) Multi-model pooling, calibrated
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Fig. S6. Same as Figure 3, but using the JRA-55 reanalysis as the reference dataset.

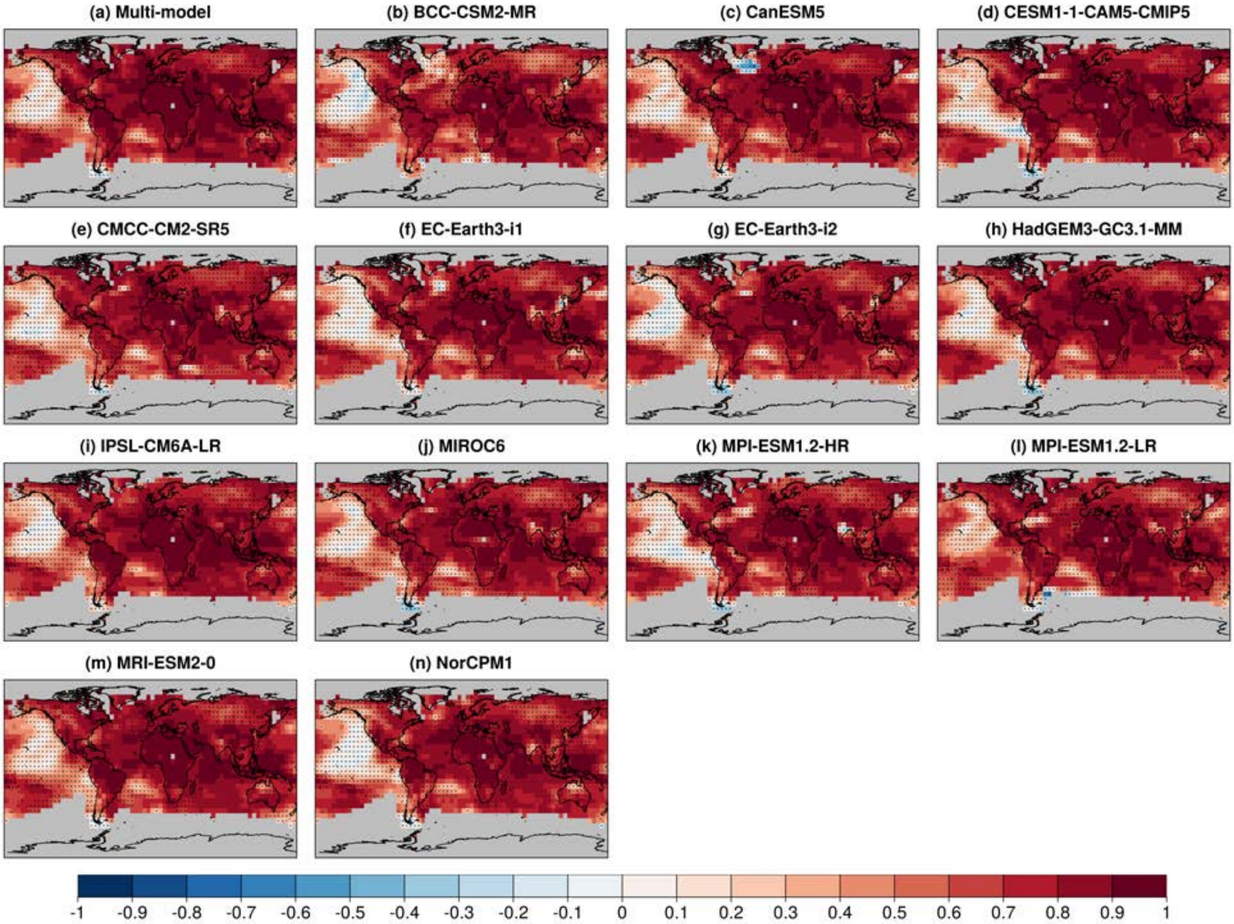


Fig. S7. ACC obtained with the multi-model (a) and the forecast systems (b–n) with decadal predictions for the forecast years 1–5 for the near-surface air temperature. The skill estimates have been computed over the 1961–2014 period (start dates 1960–2009). The reference period for the computation of anomalies is 1981–2010. The reference dataset used is the GHCNv4 dataset. Crosses indicate that the values are not statistically significant at the 95% level using a two-sided t-test accounting for autocorrelation.

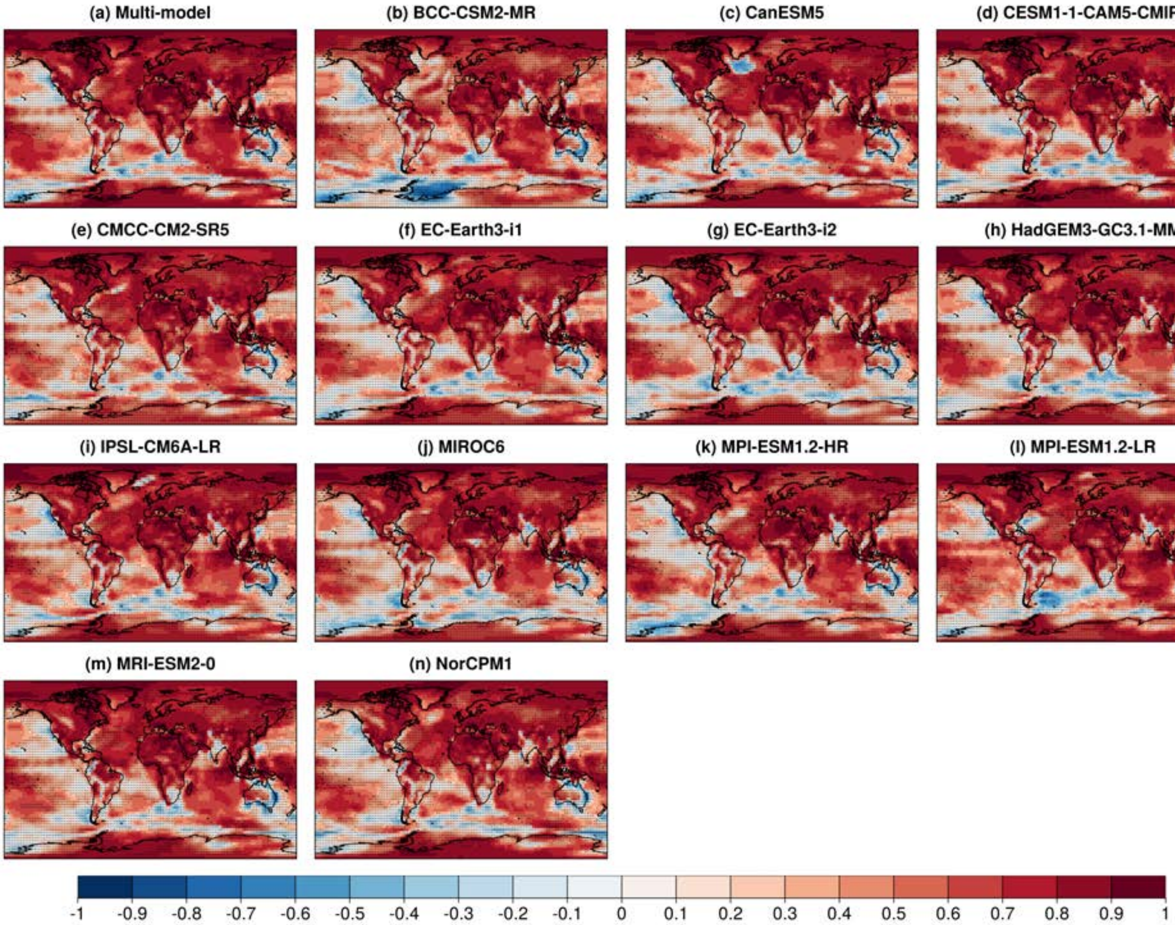


FIG. S8. Same as Figure S7, but using the JRA-55 reanalysis as the reference dataset.

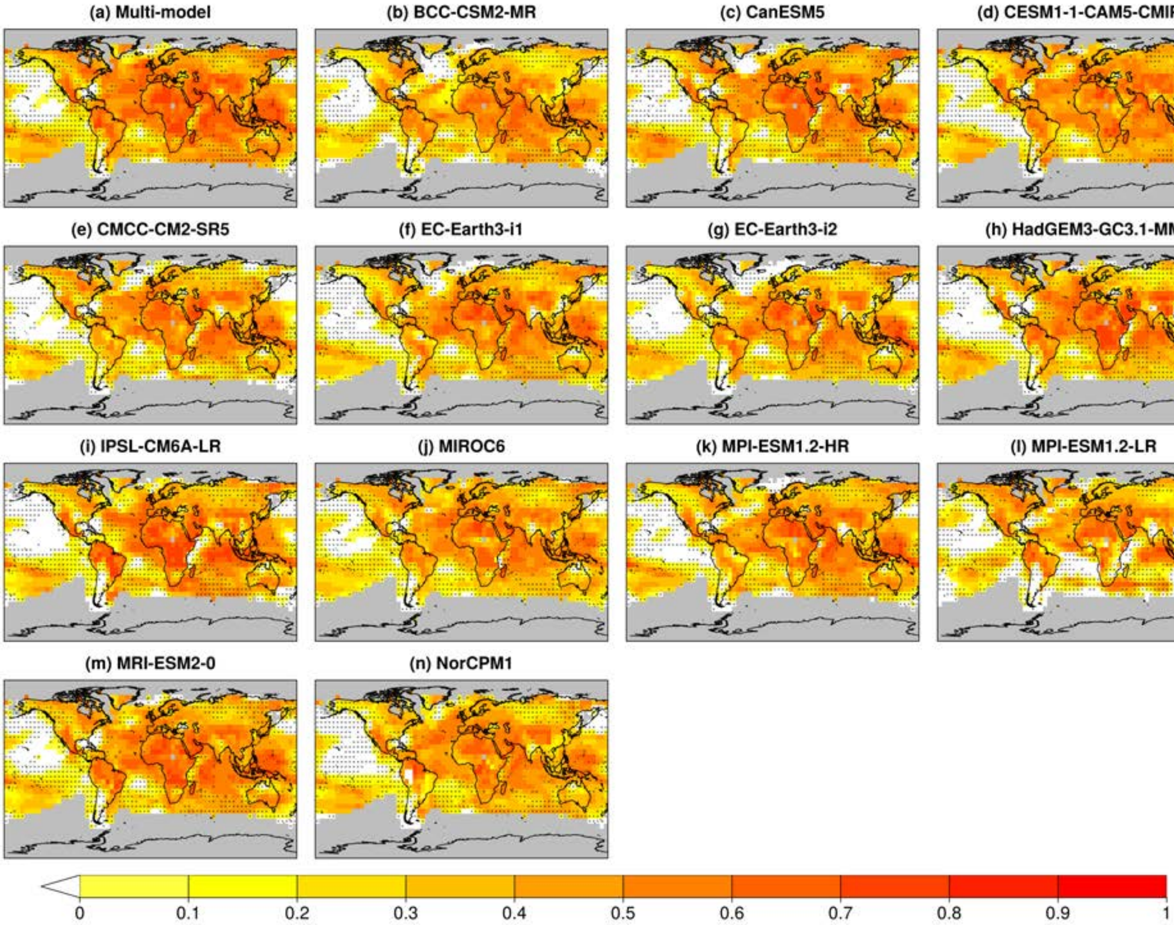


FIG. S9. RMSSS obtained with the multi-model (a) and the forecast systems (b–n) with decadal predictions for the forecast years 1–5 for the near-surface air temperature using the observed climatology as the reference forecast. The skill scores have been computed over the 1961–2014 period (start dates 1960–2009). The reference period for the computation of the anomalies is 1981–2010. The reference dataset used is the GHCNv4 dataset. Crosses indicate that the decadal predictions do not provide significantly better or worse predictions than the climatological forecast at the 95% confidence level based on a Random Walk test.

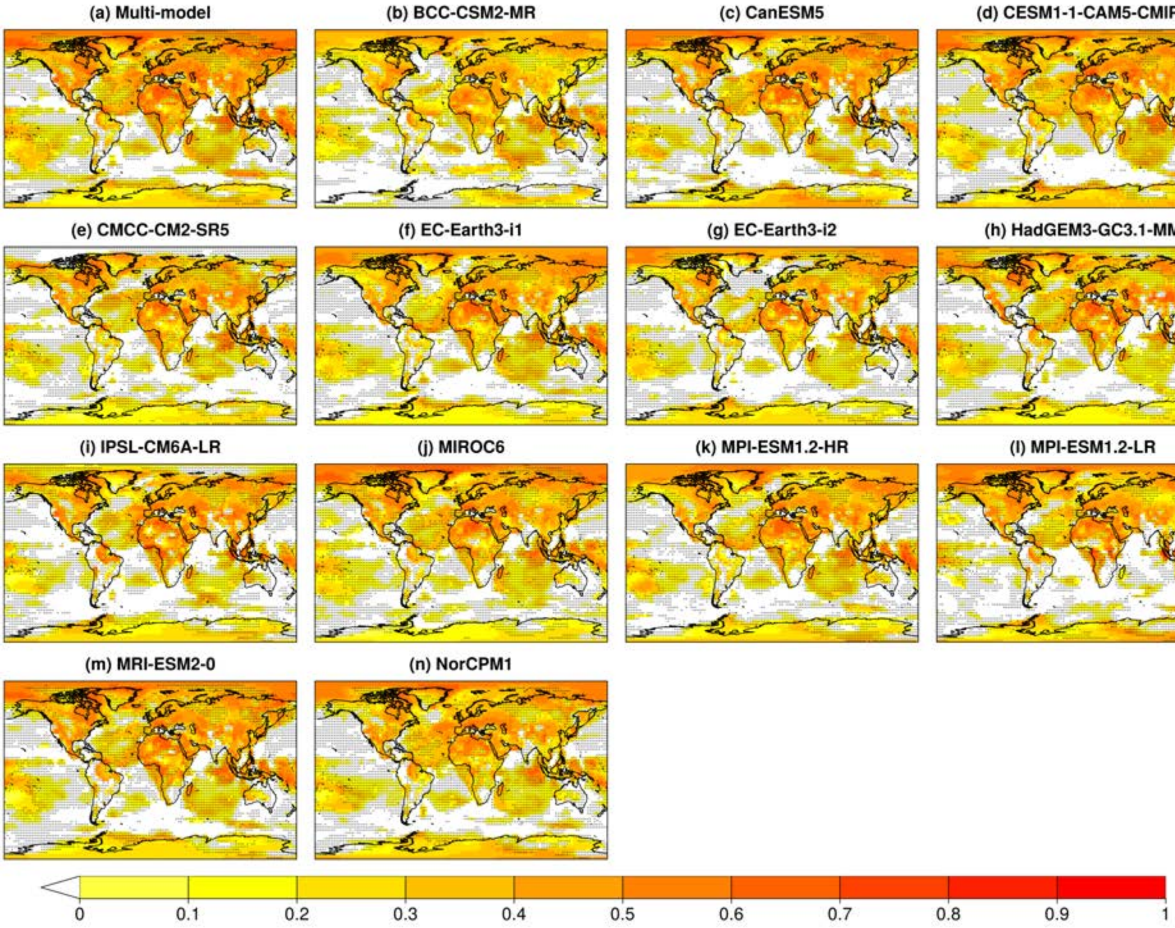


FIG. S10. Same as Figure S9, but using the JRA-55 reanalysis as the reference dataset.

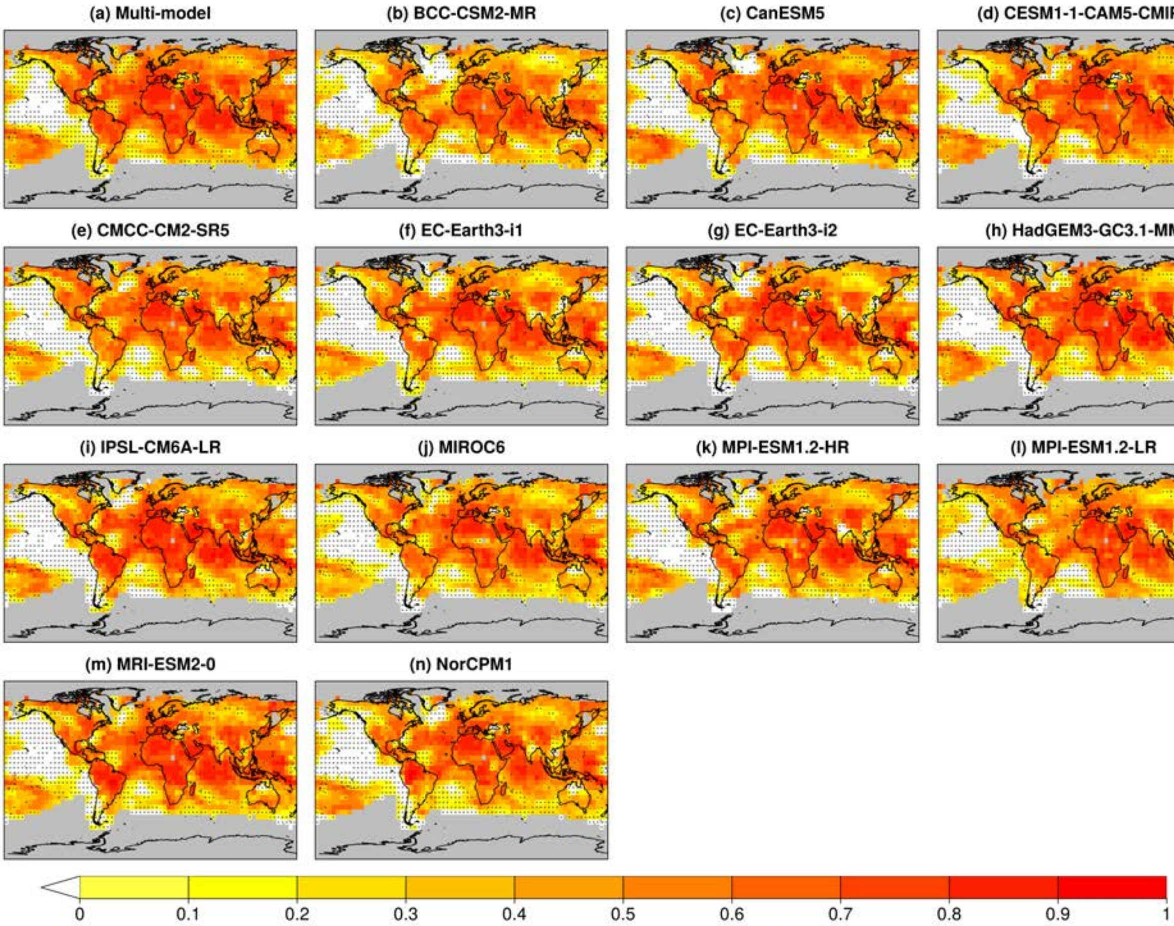


FIG. S11. RPSS for 3 categories obtained with the multi-model (a) and the forecast systems (b–n) with decadal predictions for the forecast years 1–5 for the near-surface air temperature using the observed climatology as the reference forecast. The skill scores have been computed over the 1961–2014 period (start dates 1960–2009). The reference period for the computation of the thresholds between categories is 1981–2010. The reference dataset used is the GHCNv4 dataset. Crosses indicate that the decadal predictions do not provide significantly better or worse predictions than the climatological forecast at the 95% confidence level based on a Random Walk test.

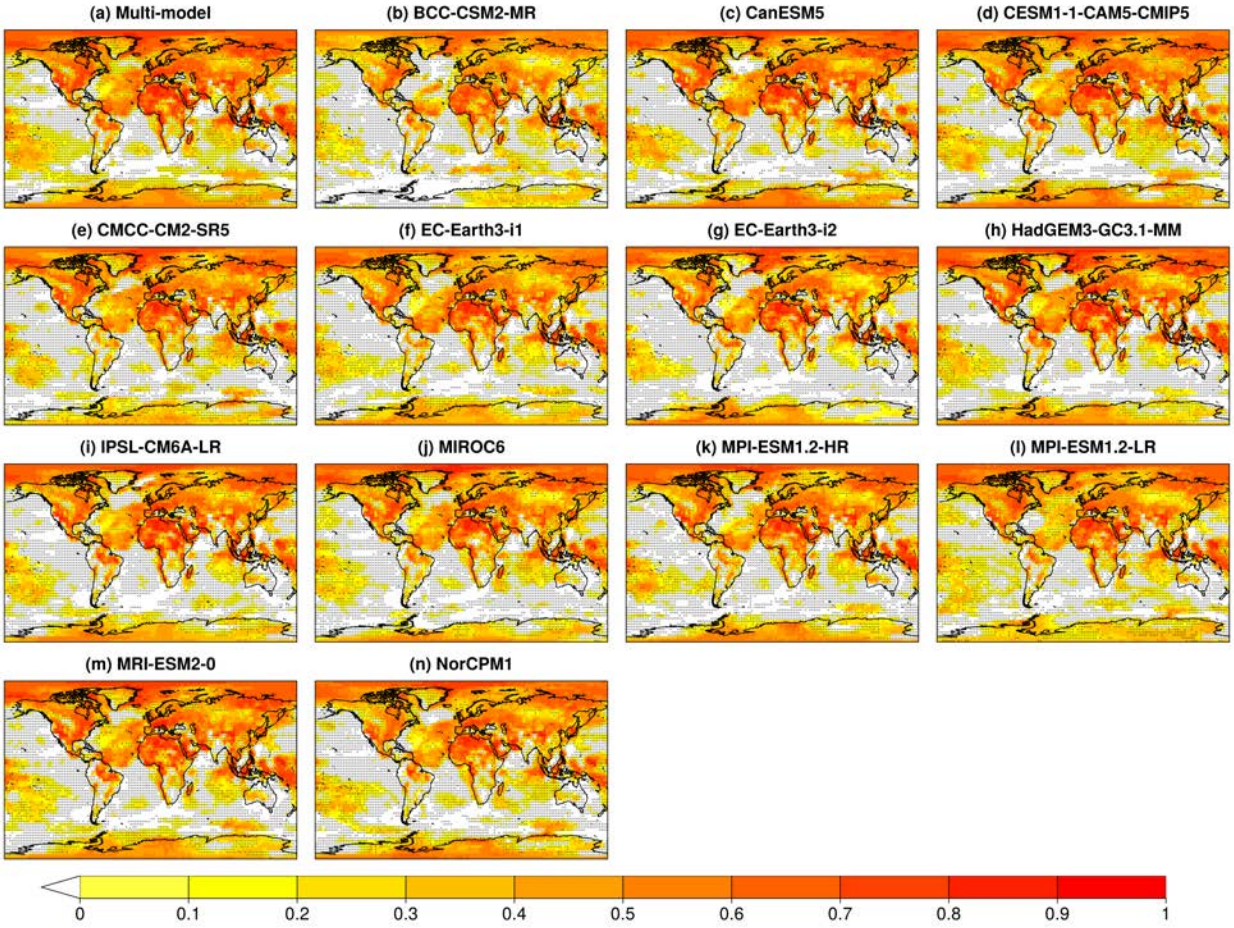


FIG. S12. Same as Figure S11, but using the JRA-55 reanalysis as the reference dataset.

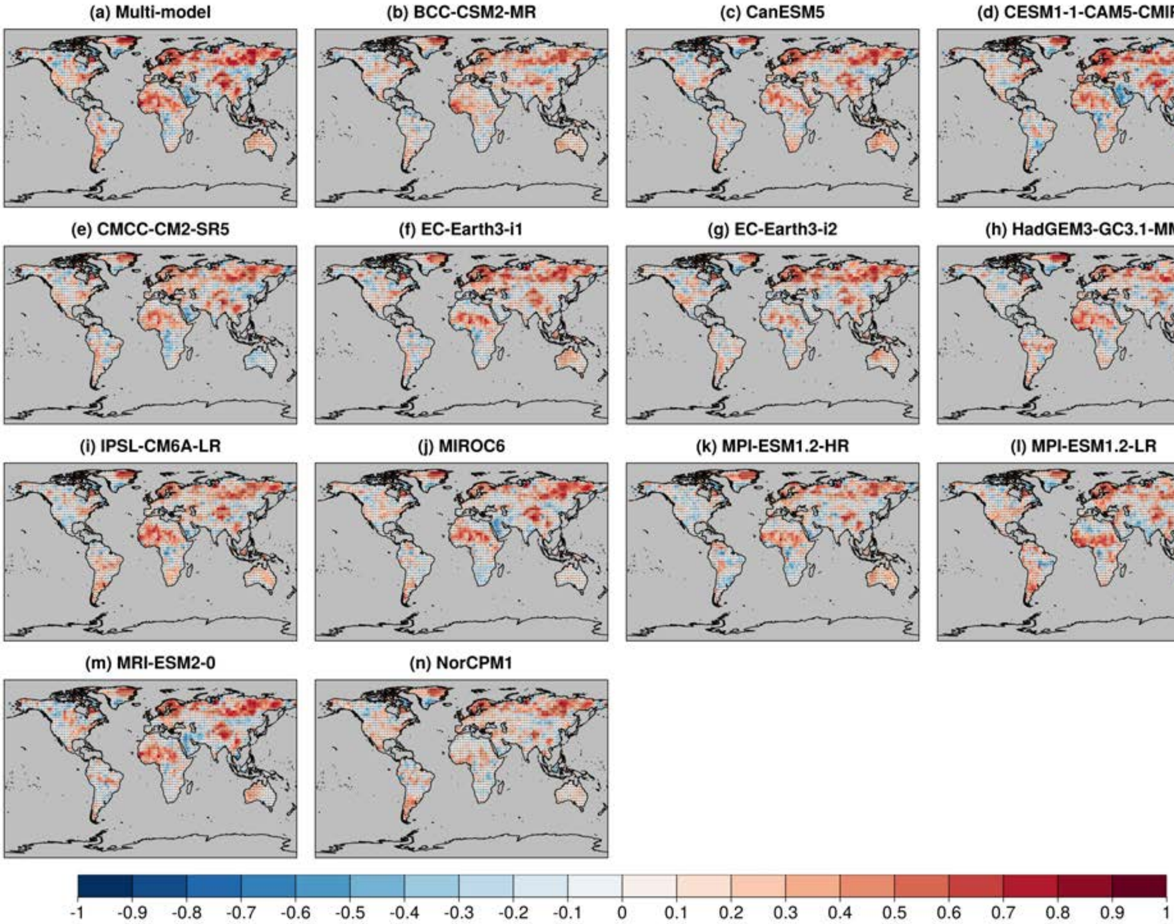


FIG. S13. ACC obtained with the multi-model (a) and the forecast systems (b–n) with decadal predictions for the forecast years 1–5 for precipitation. The skill estimates have been computed over the 1961–2014 period (start dates 1960–2009). The reference period for the computation of anomalies is 1981–2010. The reference dataset used is the GPCC dataset. Crosses indicate that the values are not statistically significant at the 95% level using a two-sided t-test accounting for autocorrelation.

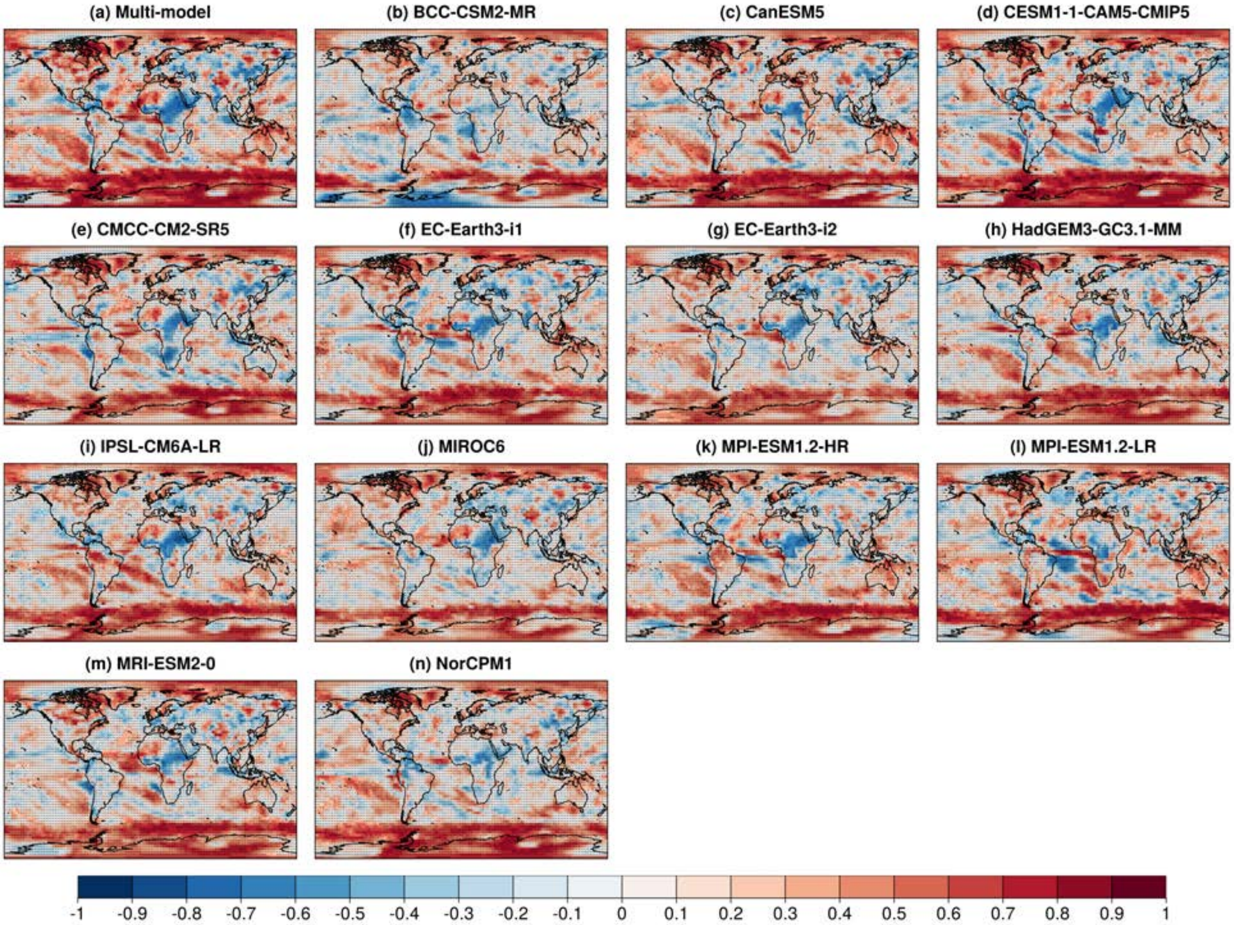


FIG. S14. Same as Figure S13, but using the JRA-55 reanalysis as the reference dataset.

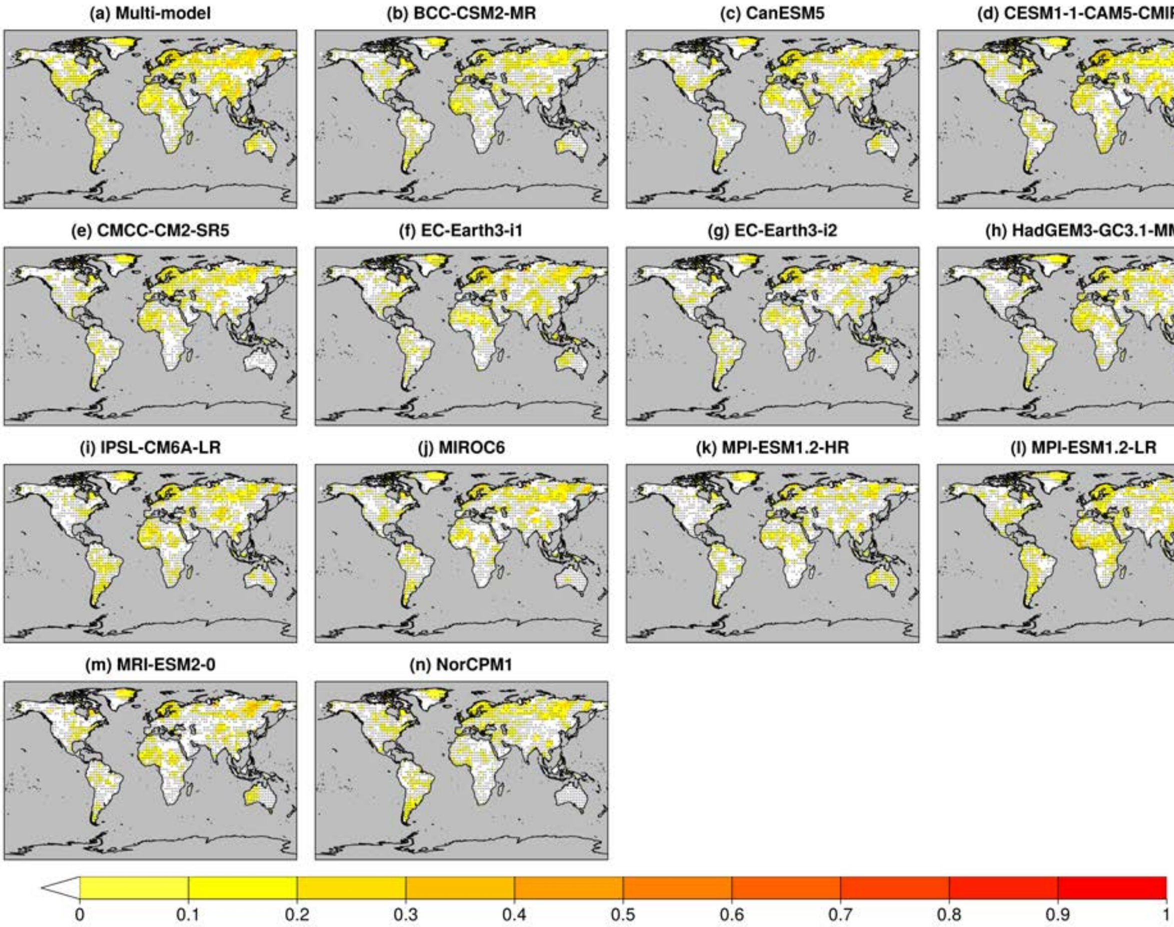


Fig. S15. RMSSS obtained with the multi-model (a) and the forecast systems (b–n) with decadal predictions for the forecast years 1–5 for precipitation using the observed climatology as the reference forecast. The skill scores have been computed over the 1961–2014 period (start dates 1960–2009). The reference period for the computation of the anomalies is 1981–2010. The reference dataset used is the GPCC dataset. Crosses indicate that the decadal predictions do not provide significantly better or worse predictions than the climatological forecast at the 95% confidence level based on a Random Walk test.

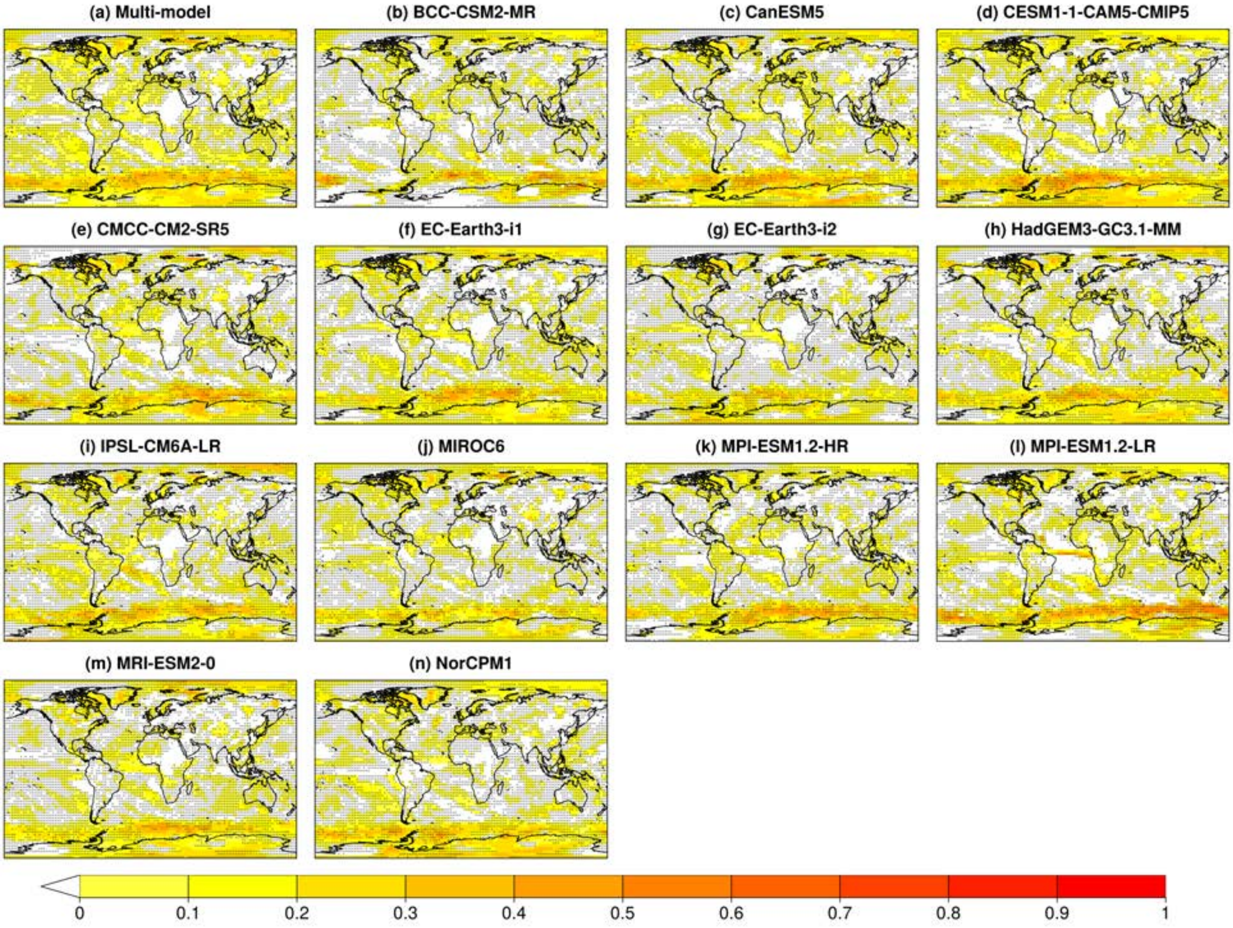


FIG. S16. Same as Figure S15, but using the JRA-55 reanalysis as the reference dataset.

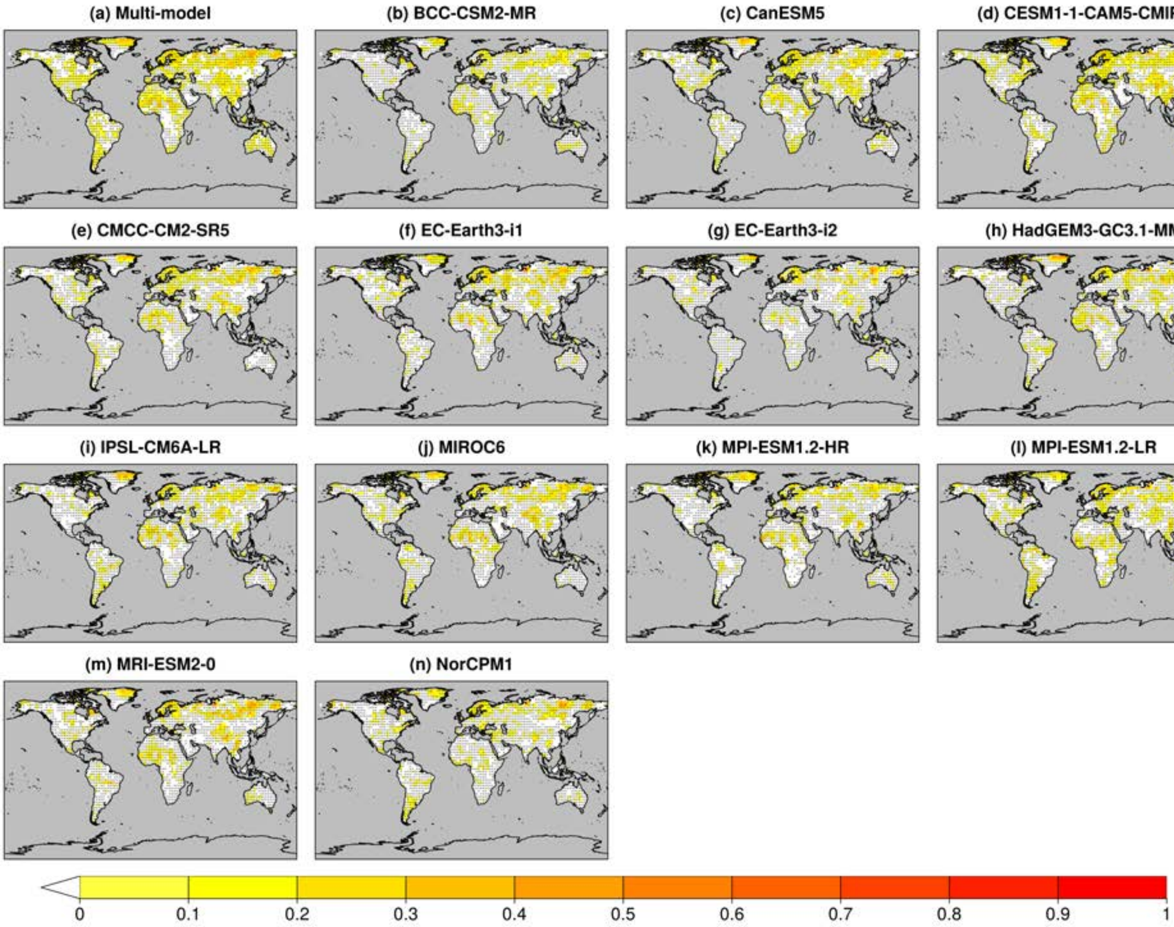


FIG. S17. RPSS for 3 categories obtained with the multi-model (a) and the forecast systems (b–n) with decadal predictions for the forecast years 1–5 for precipitation using the observed climatology as the reference forecast. The skill scores have been computed over the 1961–2014 period (start dates 1960–2009). The reference period for the computation of the thresholds between categories is 1981–2010. The reference dataset used is the GPCC dataset. Crosses indicate that the decadal predictions do not provide significantly better or worse predictions than the climatological forecast at the 95% confidence level based on a Random Walk test.

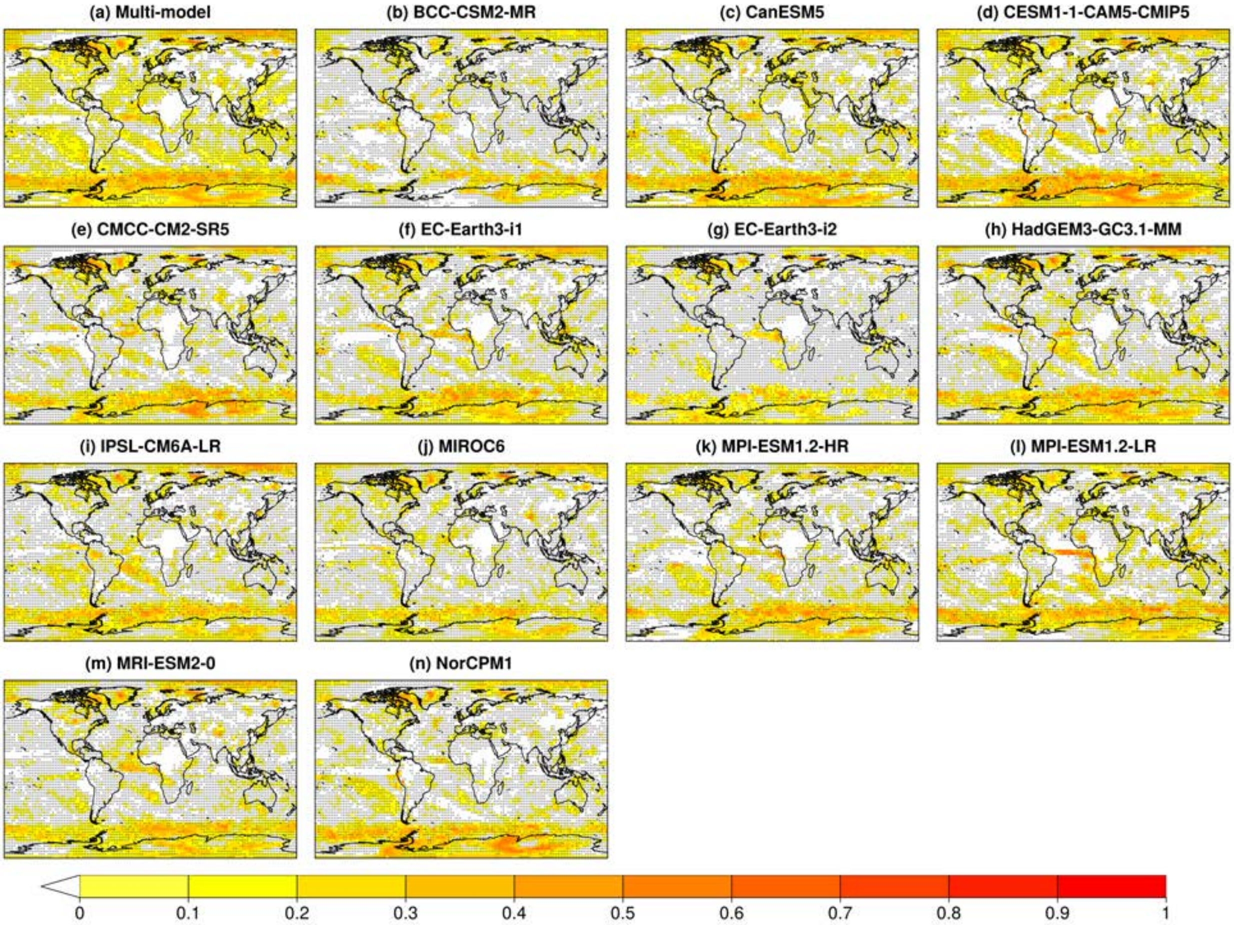


FIG. S18. Same as Figure S17, but using the JRA-55 reanalysis as the reference dataset.

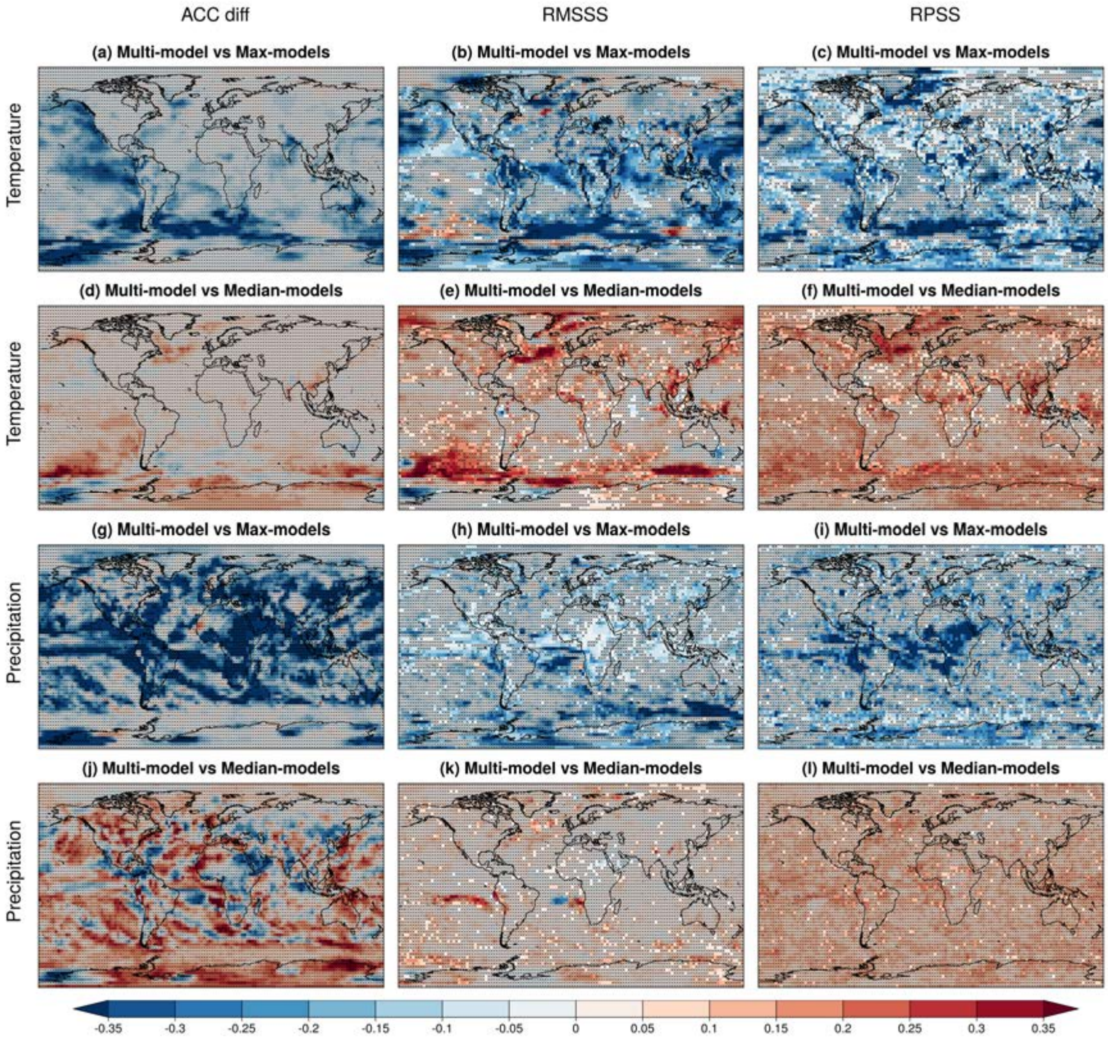

ACC diff
RMSSS
RPSS
(a) Multi-model vs Max-models
(b) Multi-model vs Max-models
(c) Multi-model vs Max-models
Temperature
(d) Multi-model vs Median-models
(e) Multi-model vs Median-models
(f) Multi-model vs Median-models
Temperature
(g) Multi-model vs Max-models
(h) Multi-model vs Max-models
(i) Multi-model vs Max-models
Precipitation
(j) Multi-model vs Median-models
(k) Multi-model vs Median-models
(l) Multi-model vs Median-models
Precipitation
-0.35
-0.3
-0.25
-0.2
-0.15
-0.1
-0.05
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35

FIG. S19. Same as Figure 5, but using the JRA-55 reanalysis as the reference dataset.

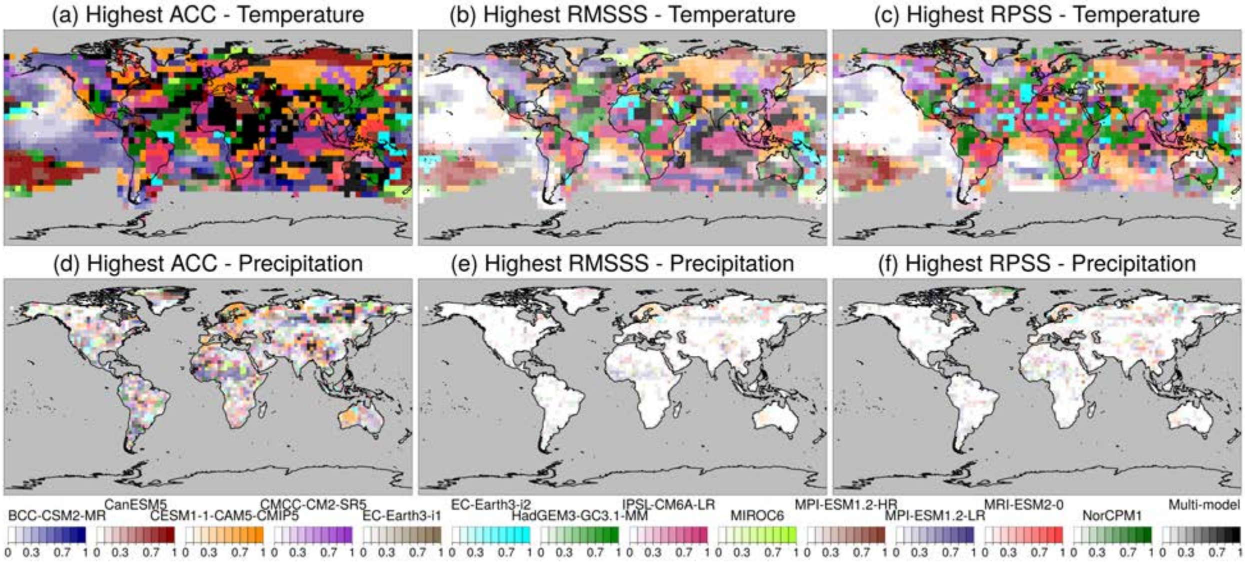

(a) Highest ACC - Temperature
(b) Highest RMSSS - Temperature
(c) Highest RPSS - Temperature
(d) Highest ACC - Precipitation
(e) Highest RMSSS - Precipitation
(f) Highest RPSS - Precipitation
BCC-CSM2-MR
CanESM5
CESM1-1-CAM5-CMIP5
CMCC-CM2-SR5
EC-Earth3-i1
EC-Earth3-i2
HadGEM3-GC3.1-MM
IPSL-CM6A-LR
MIROC6
MPI-ESM1.2-HR
MPI-ESM1.2-LR
MRI-ESM2-0
NorCPM1
Multi-model
0
0.3
0.7
1

FIG. S20. Highest ACC (a, d), RMSSS (b, e), and RPSS for three categories (c, f) obtained with the DCPP forecast systems and the multi-model for the forecast years 1–5 for the near-surface air temperature (a–c) and precipitation (d–f). The skill estimates have been computed over the 1961–2014 period (start dates 1960–2009). The reference period for the computation of anomalies and the thresholds between the categories is 1981–2010. The RMSSS and RPSS have been computed using climatology as the reference forecast. The reference datasets used for the near-surface air temperature and precipitation are, respectively, the GHCNv4 and the GPCC datasets.

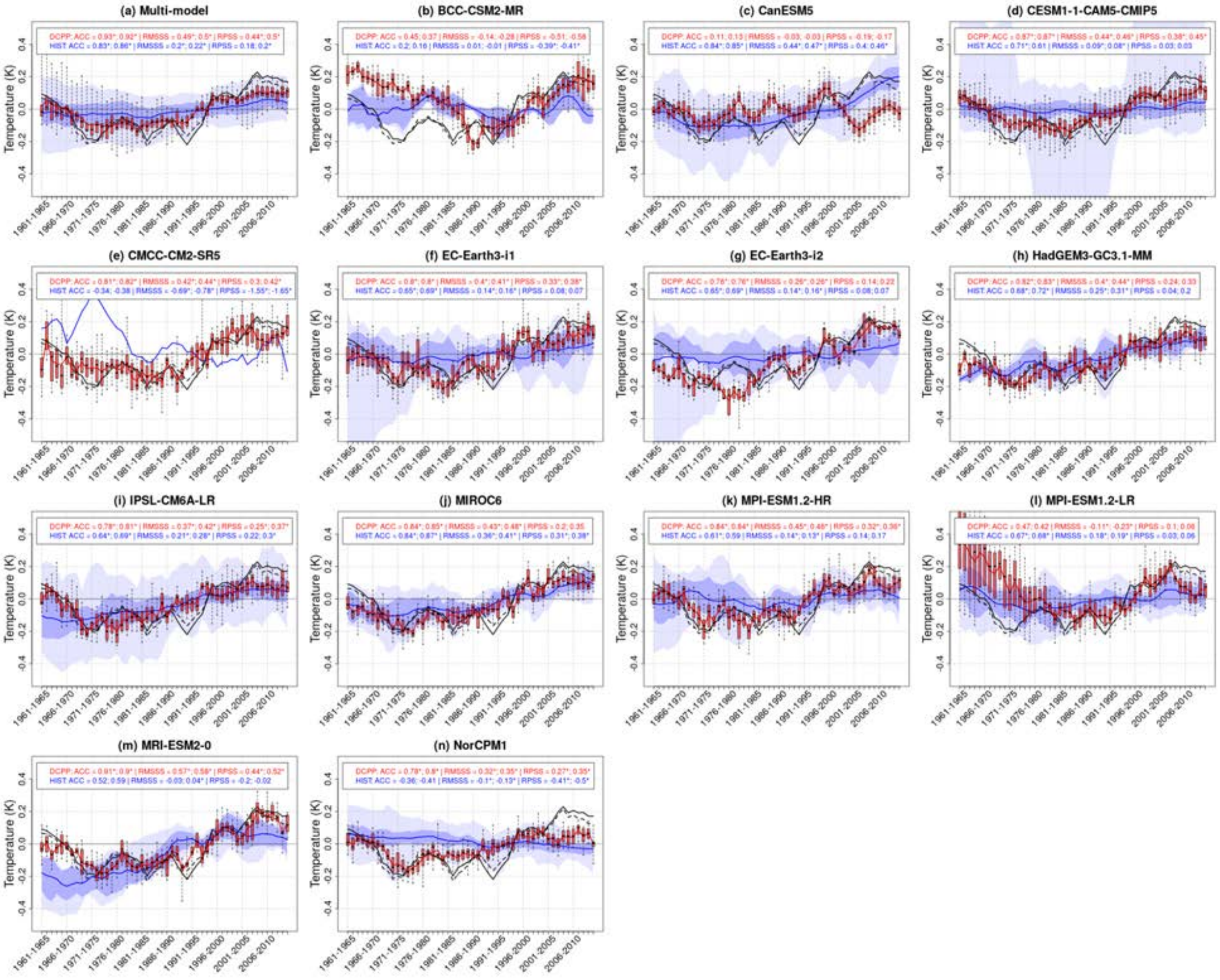


FIG. S21. AMV index obtained with the multi-model (a) and the forecast systems (b–n) for the forecast years 1–5. The historical simulations are shown in blue (dark shading contains the values between the percentiles 25th and 75th, while light shading contains the values between those percentiles and the minimum/maximum values) and the decadal predictions in red (boxes contain the values between the percentiles 25th and 75th, while the whiskers contain the values between those percentiles and the minimum/maximum values). The ACC, RMSSS, and RPSS are shown for both decadal predictions and historical simulations over 1961–2014 (start dates 1960–2009). The reference period for the computation of anomalies and thresholds between categories is 1981–2010. The reference datasets are the GISTEMPv4 (grey solid lines) and the HadCRUT4 (grey dashed lines). The skill measures are shown for both reference datasets: the first value corresponds to the GISTEMPv4 dataset and the second value to the HadCRUT4 dataset. A star next to an ACC estimate indicates that the skill is statistically significant at the 95% confidence level using a two-sided t-test accounting for autocorrelation, while a star next to an RMSSS or RPSS value indicates that the simulations provide significantly better or worse predictions than the climatological forecast at the 95% confidence level based on a Random Walk test.

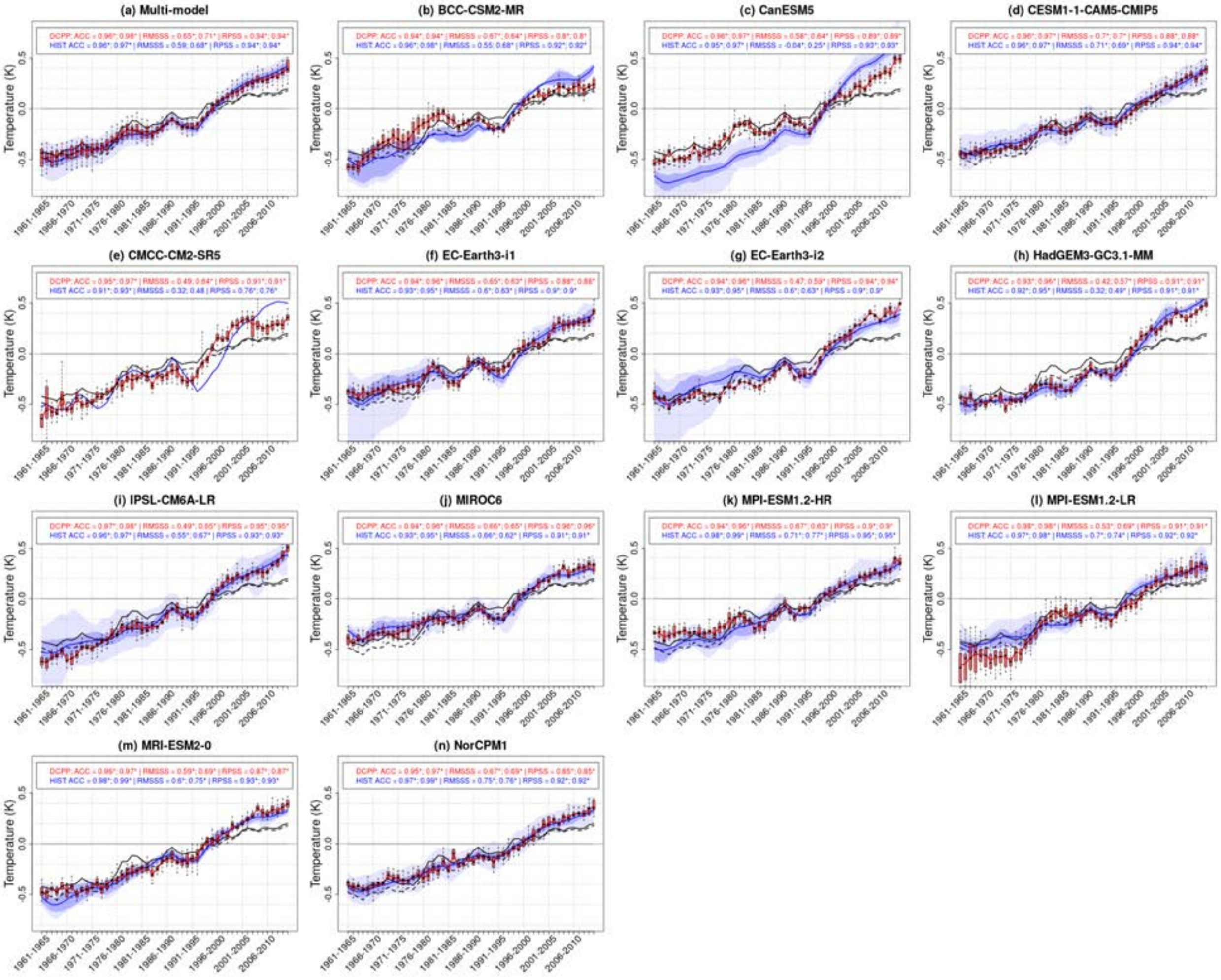


FIG. S22. GSAT anomalies obtained with the multi-model (a) and the forecast systems (b-n) for the forecast years 1–5. The historical simulations are shown in blue (ensemble spread shown by shading) and the decadal predictions in red (ensemble spread shown by box-and-whiskers). The ACC, RMSSS, and RPSS are shown for both decadal predictions and historical simulations over 1961–2014 (start dates 1960–2009). The reference period for the computation of anomalies and thresholds between categories is 1981–2010. The reference datasets are the JRA-55 (grey solid lines) and the GHCNv4 (grey dashed lines). The skill measures are shown for both reference datasets: the first value corresponds to the JRA-55 dataset and the second value to the GHCNv4 dataset. A star next to an ACC estimate indicates that the skill is statistically significant at the 95% confidence level using a two-sided t-test accounting for autocorrelation, while a star next to an RMSSS or RPSS value indicates that the simulations provide significantly better or worse predictions than the climatological forecast at the 95% confidence level based on a Random Walk test.

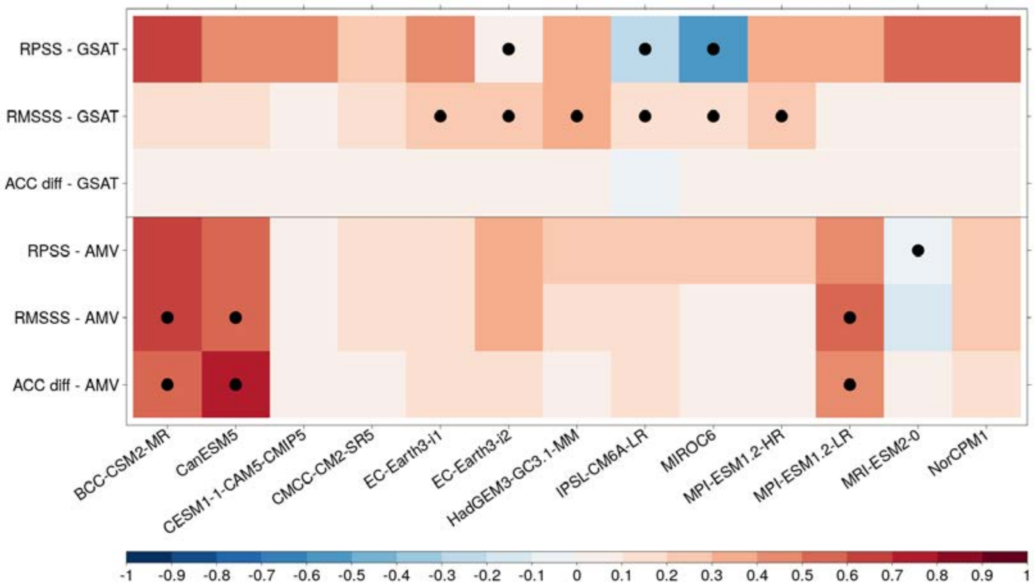


Fig. S23. Same as Figure 6, but using the HadCRUT4 and GHCNv4 datasets for the AMV and GSAT indices, respectively, as the reference datasets.

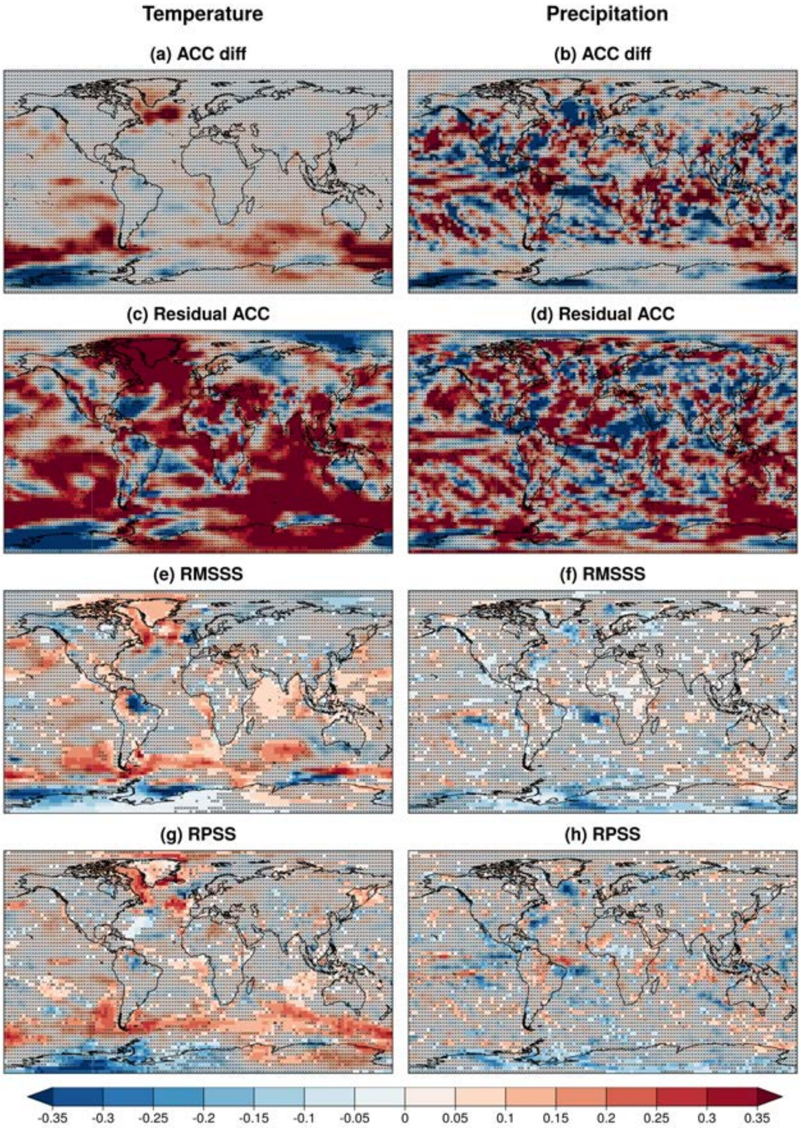


FIG. S24. Same as Figure 7, but using the JRA-55 reanalysis as the reference dataset.

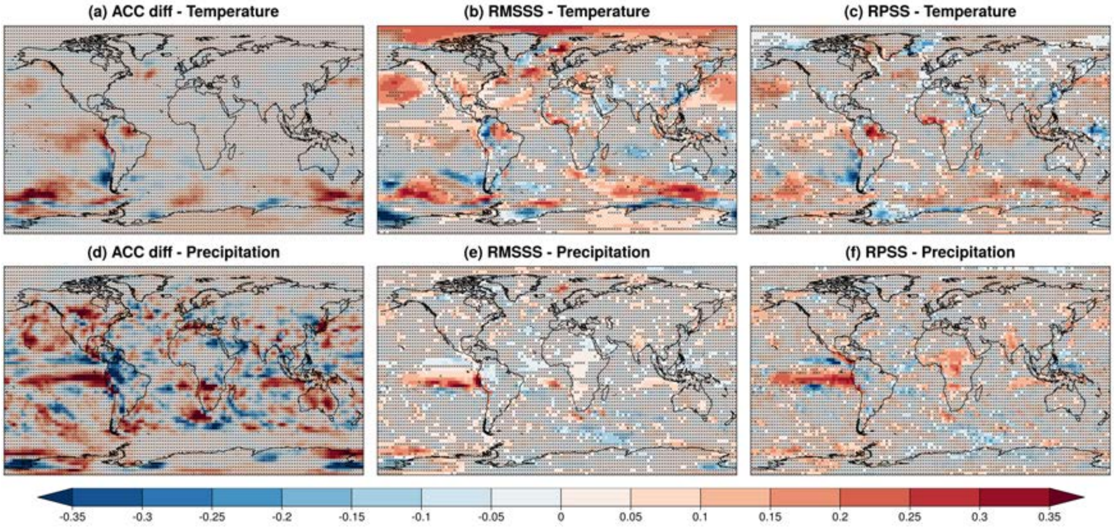


FIG. S25. Same as Figure 8, but using the JRA-55 reanalysis as the reference dataset.

Appendix C

Supplementary material for Chapter 4

Delgado-Torres, C., Verfaillie, D., Mohino, E., and Donat, M. G. (2022). Representation and annual to decadal predictability of Euro-Atlantic weather regimes in the CMIP6 version of the EC-Earth coupled climate model. Journal of Geophysical Research: Atmospheres, 127, e2022JD036673. https://doi.org/10.1029/2022JD036673

Main objectives, main outcomes and research article in Chapter 4.

Supporting Information for ”Representation and annual to decadal predictability of Euro-Atlantic weather regimes in the CMIP6 version of the EC-Earth coupled climate model”

DOI: 10.1029/2022JD036673

C. Delgado-Torres[1,2], D. Verfaillie[1,3], E. Mohino[2], M. G. Donat[1,4]

[1]Barcelona Supercomputing Center (BSC), Barcelona, Spain

[2]Department of Physics of the Earth and Astrophysics, Universidad Complutense de Madrid, Madrid, Spain

[3]Aix-Marseille University, CNRS, IRD, Coll. France, INRAE, CEREGE, Aix-en-Provence, France

[4]Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluis Companys 23, Barcelona, Spain

Contents of this file

1. Figures S1 to S8

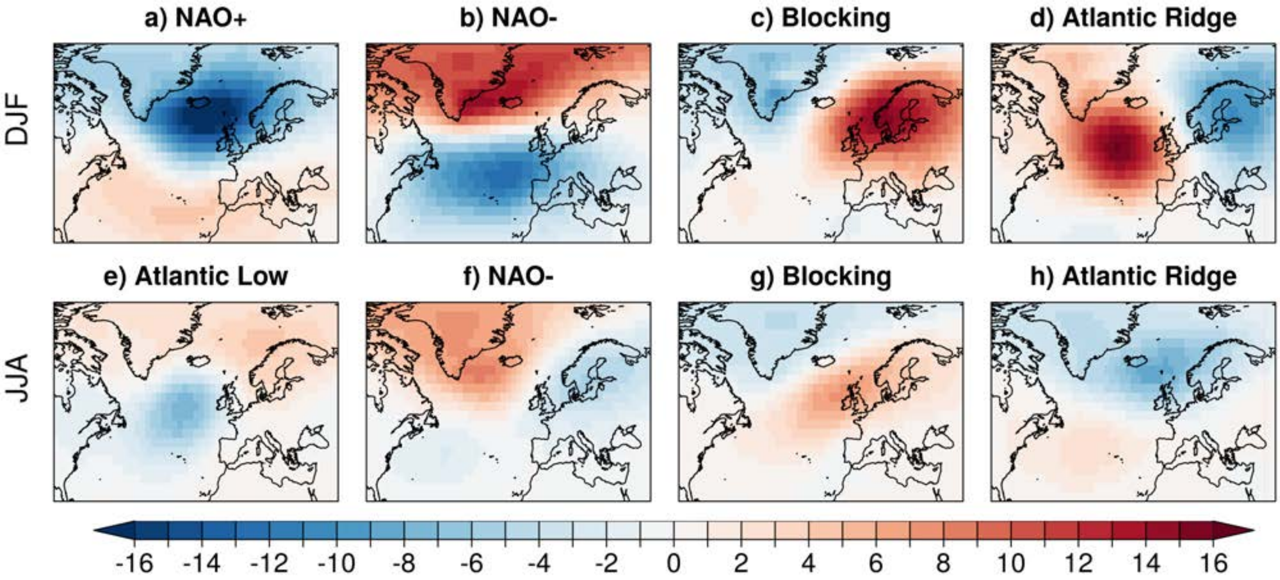


Figure S1. Same as Figure 1, but using the NCEP1 reanalysis as the reference dataset.

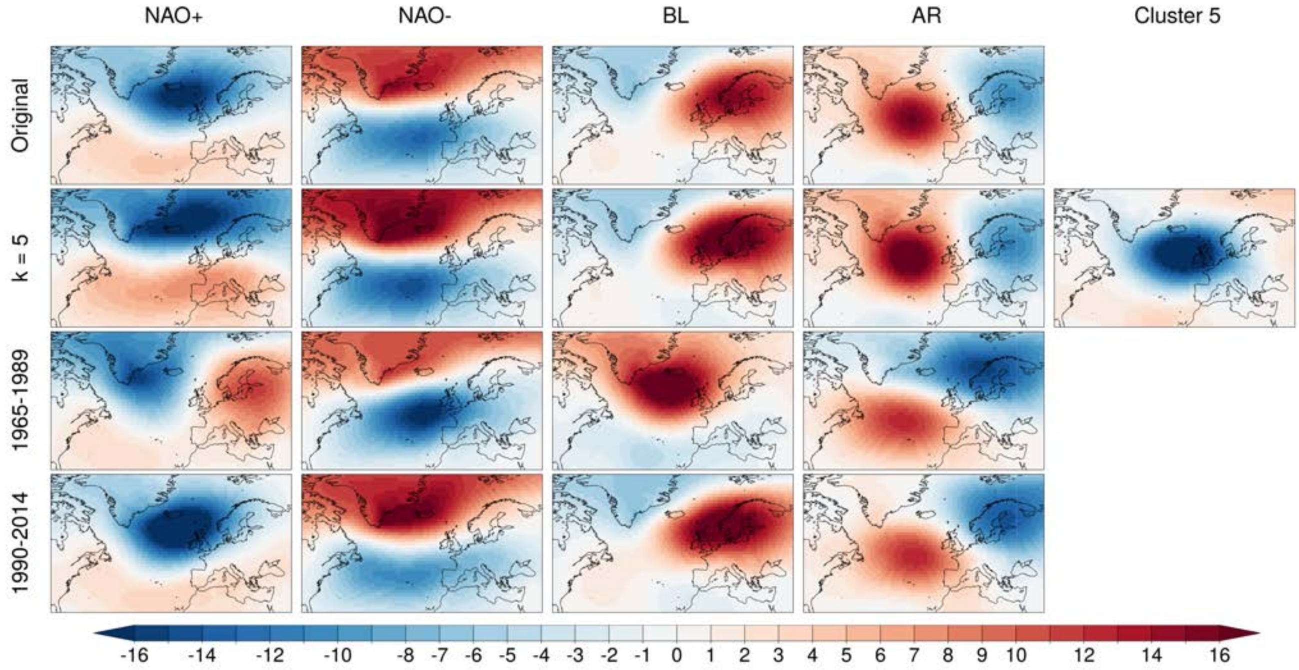


Figure S2. Spatial patterns of the observed Euro-Atlantic weather regimes (computed as the averaged sea level pressure anomalies, in hPa, of all the days classified onto each cluster) obtained with the JRA-55 reanalysis for the winter season. The first and second rows show the patterns obtained by applying the k-means clustering asking for 4 and 5 clusters, respectively, during the 1965–2014 period. The third and fourth rows show the patterns obtained by applying the k-means clustering algorithm asking for 4 clusters during the 1995–1989 and 1990-2014 periods, respectively.

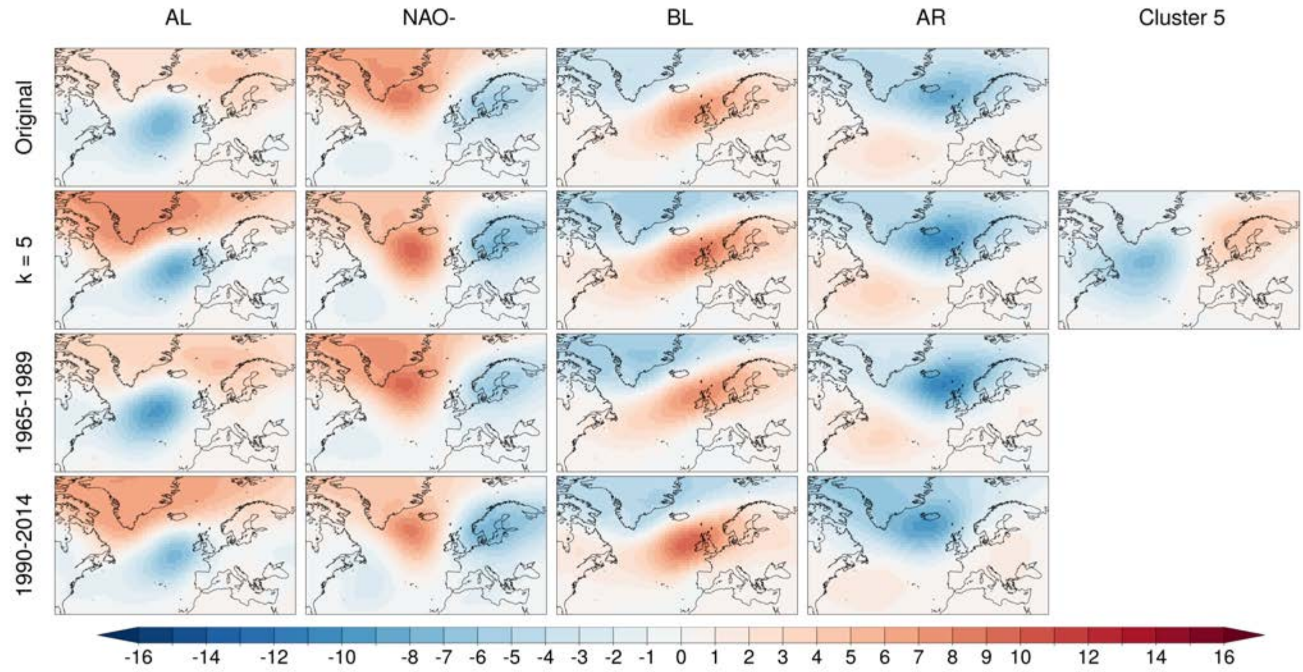


Figure S3. Same as Figure S2, but for the summer season.

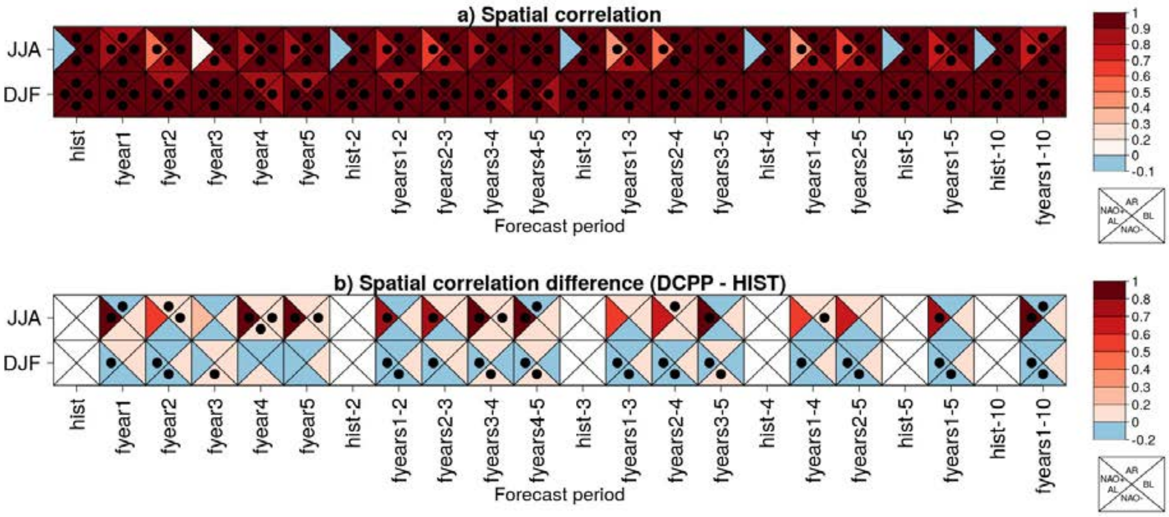


Figure S4. Same as Figure 2, but using the NCEP1 reanalysis as the reference dataset.

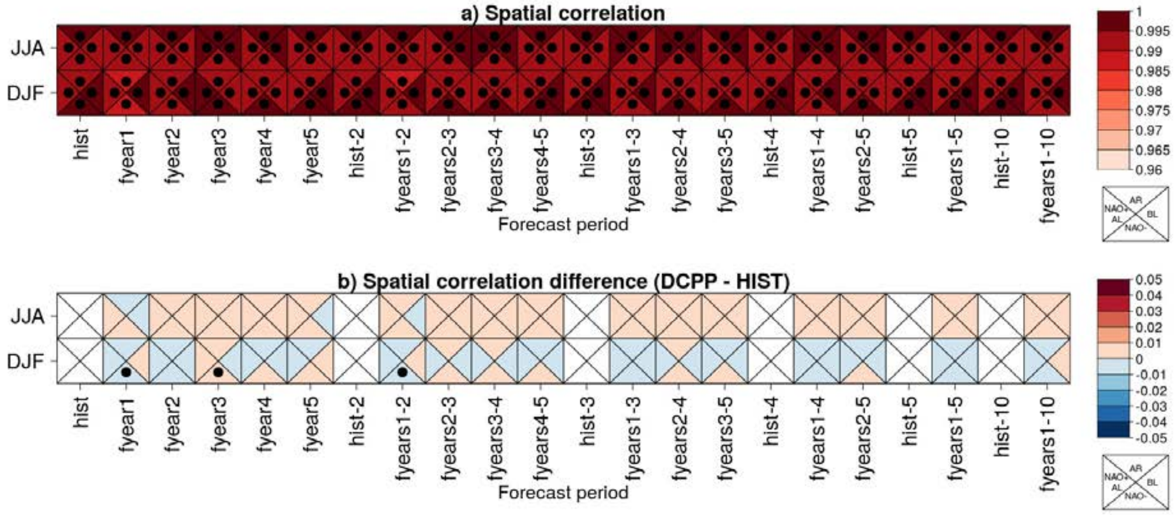


Figure S5. Same as Figure 2, but projecting the simulated daily maps onto the observed clusters during the weather regimes computation.

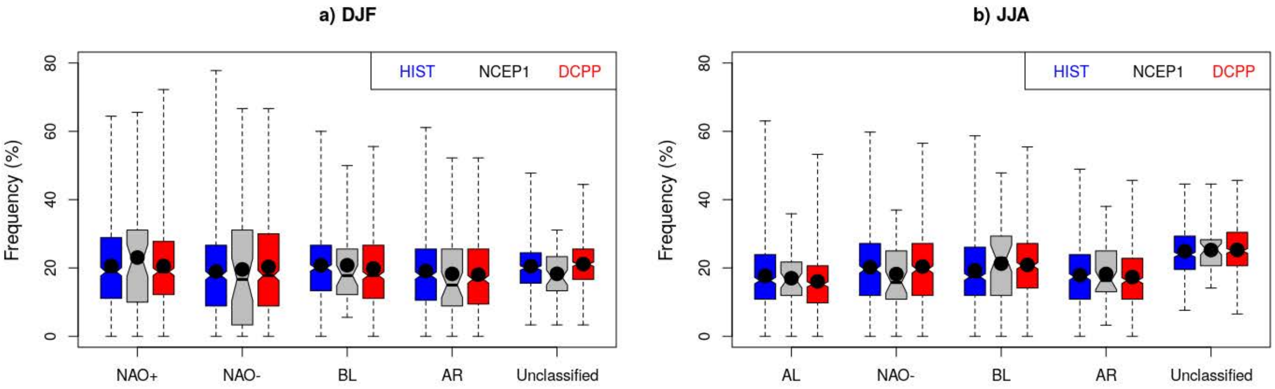


Figure S6. Same as Figure 3, but using the NCEP1 reanalysis as the reference dataset.

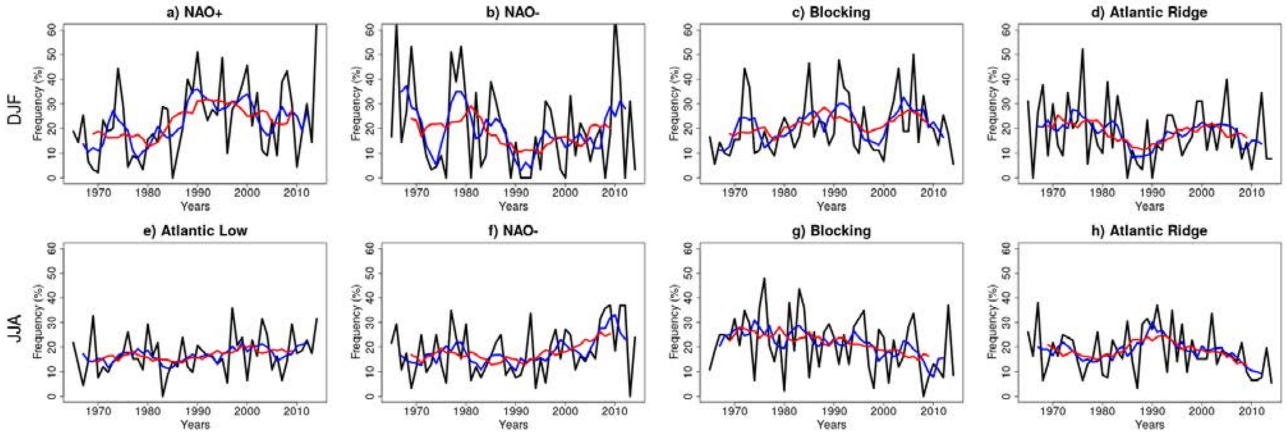


Figure S7. Same as Figure 4, but using the NCEP1 reanalysis as the reference dataset.

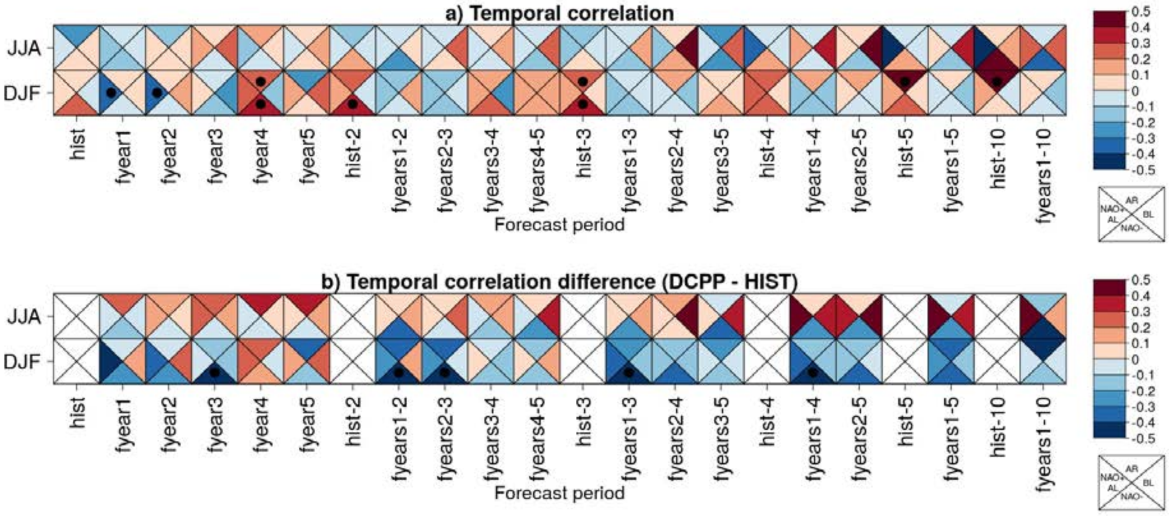


Figure S8. Same as Figure 5, but using the NCEP1 reanalysis as the reference dataset.

Appendix D

Supplementary material for Chapter 5

Delgado-Torres, C., Donat, M. G., Soret, A., Gonzalez-Reviriego, N., Bretonnière, P.-A., Ho, A.-C., Pérez-Zanón, N., Samsó Cabré, M., and Doblas-Reyes, F. J. (2023). Multi-annual predictions of the frequency and intensity of daily temperature and precipitation extremes. Environmental Research Letters, 18 034031. https://doi.org/10.1088/1748-9326/acbbe1

Main objectives, main outcomes and research article in Chapter 5.

Supplementary material for ”Multi-annual predictions of the frequency and intensity of daily temperature and precipitation extremes”

Carlos Delgado-Torres[1], Markus G. Donat[1,2], Albert Soret[1], Nube González-Reviriego[1], Pierre-Antoine Bretonnière[1], An-Chi Ho[1], Núria Pérez-Zanón[1], Margarida Samsó Cabré[1], and Francisco J. Doblas-Reyes[1,2]

[1]Barcelona Supercomputing Center (BSC), Barcelona, Spain

[2]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

Table S1. Forecast systems, institution, number of decadal prediction and historical simulation members, spatial resolution of the atmospheric model, month of initialization, and reference in which they are described.

Forecast system	DCPP members	HIST members	Spatial resolution	Month initialization	Reference
BCC-CSM2-MR	8	2	1.125º x 1.125º	January	Wu et al. (2019)
CanESM5	20	25	2.8º x 2.8º	January	Swart et al. (2019)
CMCC-CM2-SR5	10	6	0.9º x 1.25º	November	Nicolì et al. (2023)
EC-Earth3-i1	10	10	0.7º x 0.7º	November	Döscher et al. (2022)
EC-Earth3-i2	5	–	0.7º x 0.7º	November	Döscher et al. (2022)
EC-Earth3-i4	10	–	0.7º x 0.7º	November	Döscher et al. (2022)
HadGEM3-GC3.1-MM	10	4	0.55º x 0.83º	November	Sellar et al. (2020)
IPSL-CM6A-LR	10	31	1.25º x 2.5º	January	Boucher et al. (2020)
MIROC6	10	10	1.4º x 1.4º	November	Tatebe et al. (2019)
MPI-ESM1.2-HR	10	10	0.9º x 0.9º	November	Müller et al. (2018)
MRI-ESM2-0	10	6	1.125º x 1.125º	November	Yukimoto et al. (2019)
NorCPM1-i1	10	30	1.9º x 2.5º	October	Bethke et al. (2021)
NorCPM1-i2	10	–	1.9º x 2.5º	October	Bethke et al. (2021)

Table S2. Reference datasets used for the evaluation, variable, institution, temporal frequency, dataset type, spatial resolution, and the reference in which they are described.

Variable	Dataset	Temporal frequency	Type	Spatial resolution	Reference
Near-surface air maximum and minimum temperature	BEST	Daily	Gridded observations	1º x 1º	–
Precipitation	REGEN	Daily	Gridded observations	1º x 1º	Contractor et al. (2020)
ETCCDI indices	HadEX3	Daily	Gridded indices	1.875º x 1.25º	Dunn et al. (2020)
Near-surface air temperature	GHCNv4	Monthly	Gridded observations	5º x 5º	Menne et al. (2018)
Precipitation	GPCC	Monthly	Gridded observations	1º x 1º	Schneider et al. (2020)

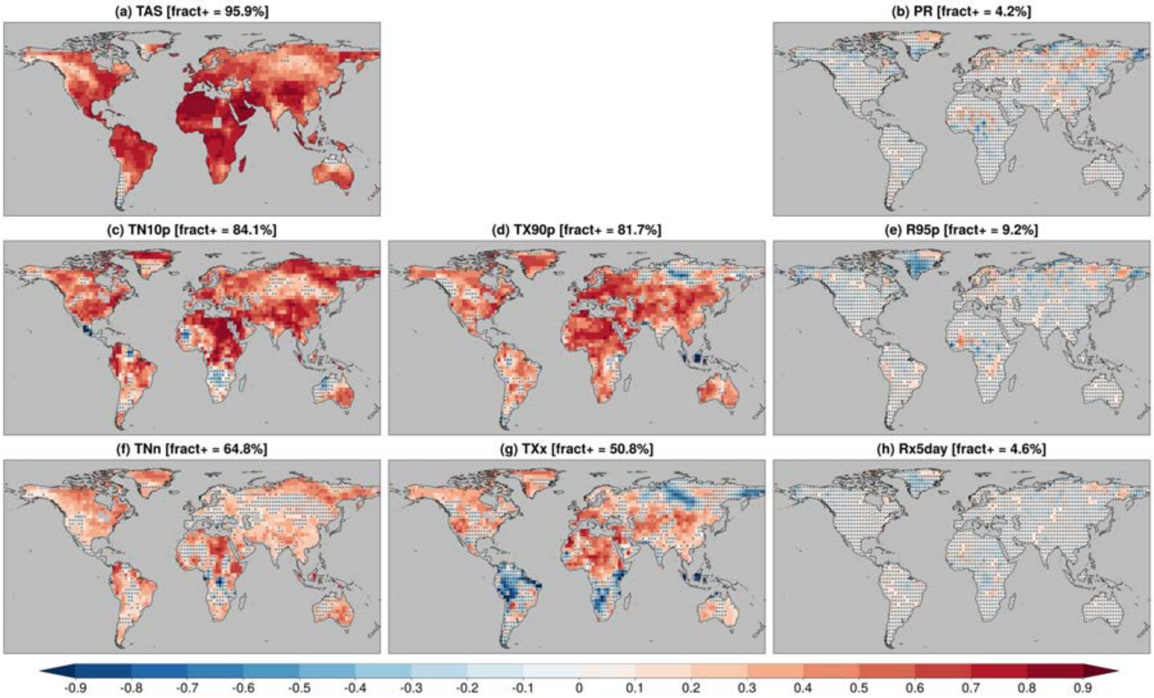


Figure S1. As Figure 1, but for the RPSS of the DCPP multi-model ensemble using the climatological forecast as the reference forecast. The statistical significance has been estimated with the Random Walk test controlling the FDR with $\alpha_{FDR} = 0.1$.

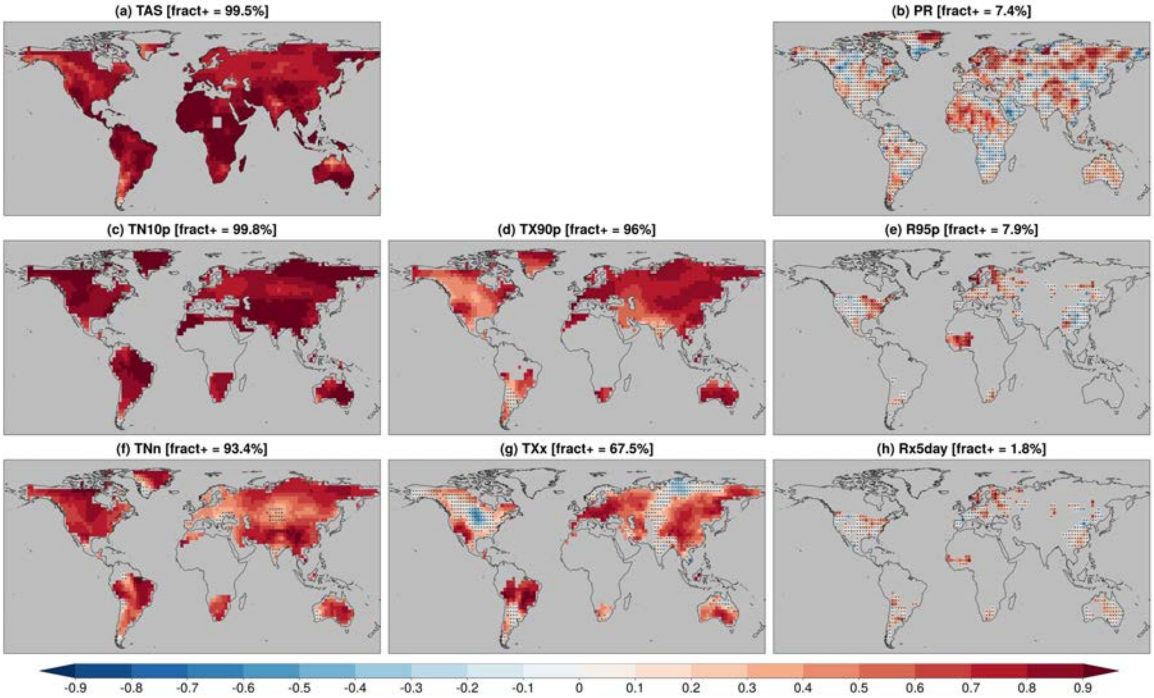


Figure S2. As Figure 1, but using the HadEX3 as the reference dataset for the extremes indices.

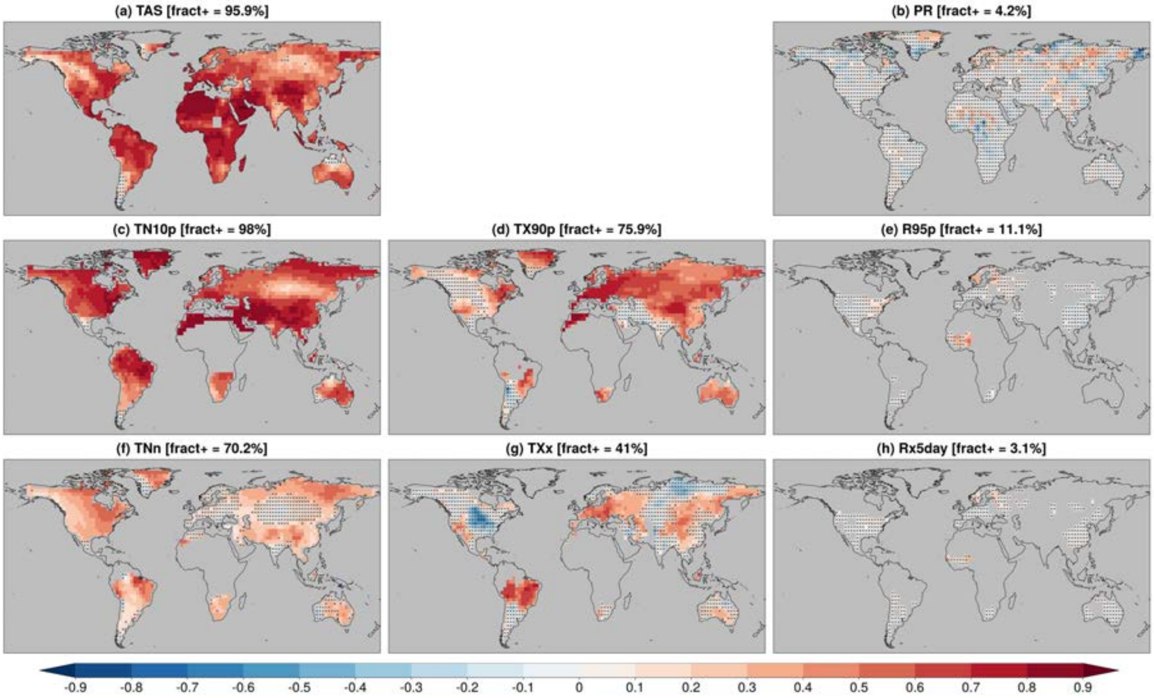


Figure S3. As Figure S1, but using HadEX3 as the reference dataset for the extreme indices.

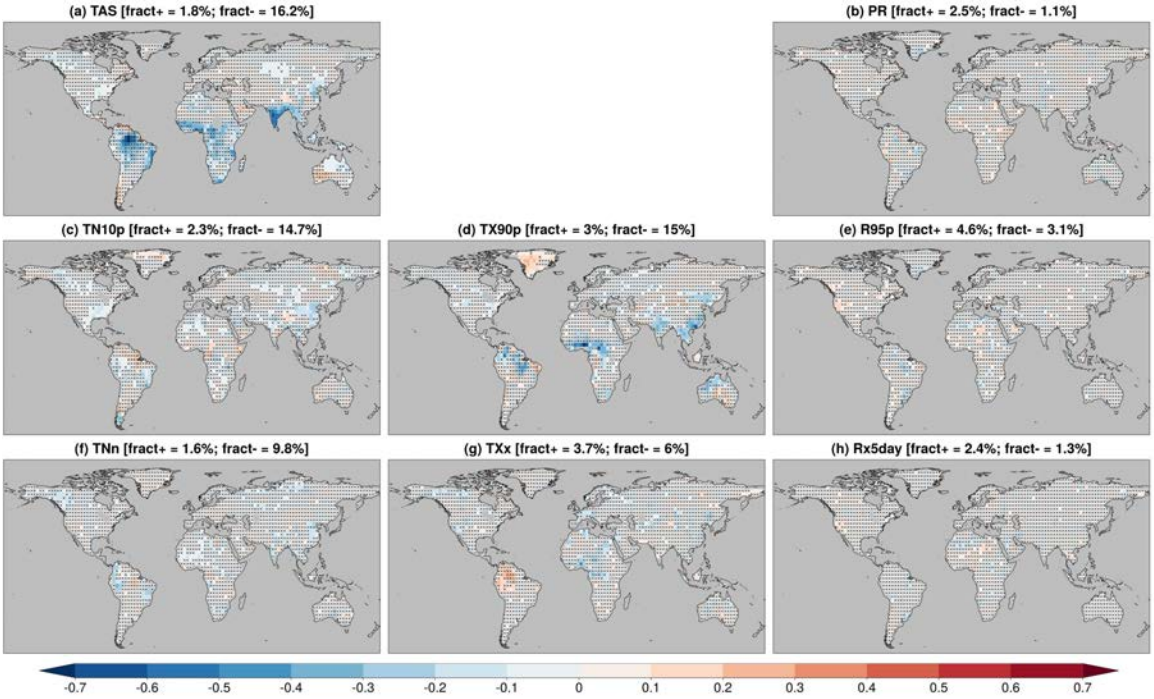

(a) TAS [fract+ = 1.8%; fract- = 16.2%]
(b) PR [fract+ = 2.5%; fract- = 1.1%]
(c) TN10p [fract+ = 2.3%; fract- = 14.7%]
(d) TX90p [fract+ = 3%; fract- = 15%]
(e) R95p [fract+ = 4.6%; fract- = 3.1%]
(f) TNn [fract+ = 1.6%; fract- = 9.8%]
(g) TXx [fract+ = 3.7%; fract- = 6%]
(h) Rx5day [fract+ = 2.4%; fract- = 1.3%]
-0.7
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

Figure S4. As Figure 2, but for the RPSS of the DCPP multi-model ensemble using the HIST multi-model ensemble as the reference forecast. The statistical significance has been estimated with the Random Walk test controlling the FDR with $\alpha_{FDR} = 0.1$.

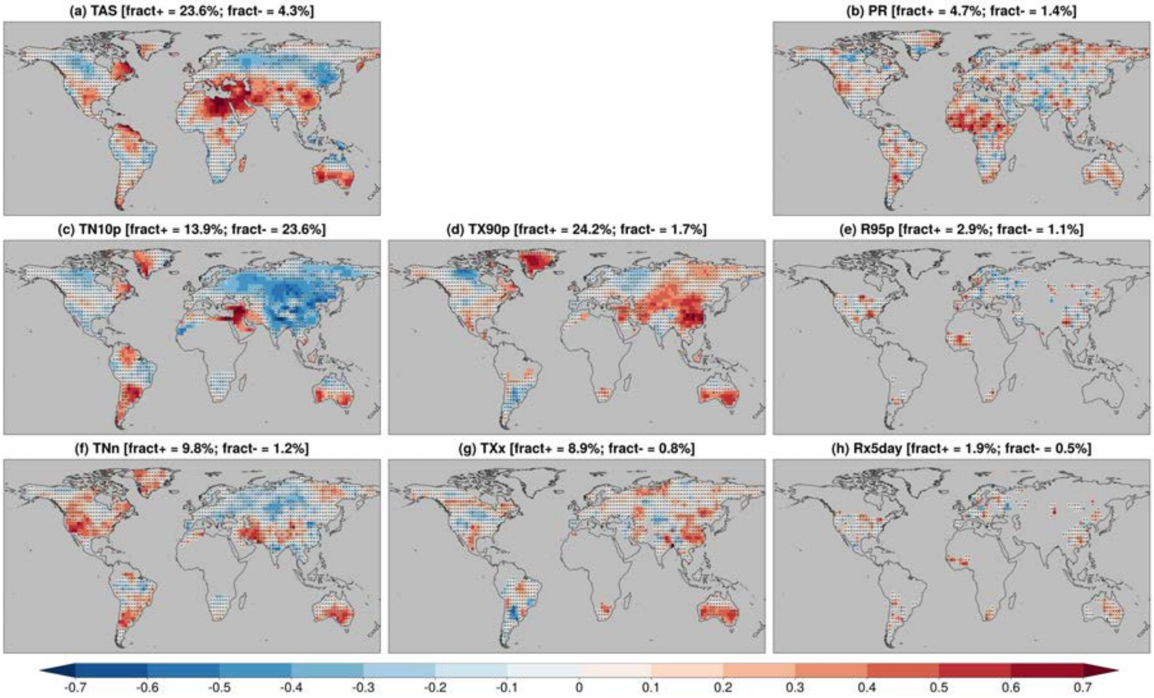

(a) TAS [fract+ = 23.6%; fract- = 4.3%]
(b) PR [fract+ = 4.7%; fract- = 1.4%]
(c) TN10p [fract+ = 13.9%; fract- = 23.6%]
(d) TX90p [fract+ = 24.2%; fract- = 1.7%]
(e) R95p [fract+ = 2.9%; fract- = 1.1%]
(f) TNn [fract+ = 9.8%; fract- = 1.2%]
(g) TXx [fract+ = 8.9%; fract- = 0.8%]
(h) Rx5day [fract+ = 1.9%; fract- = 0.5%]
-0.7
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

Figure S5. As Figure 2, but using HadEX3 as the reference dataset for the extreme indices.

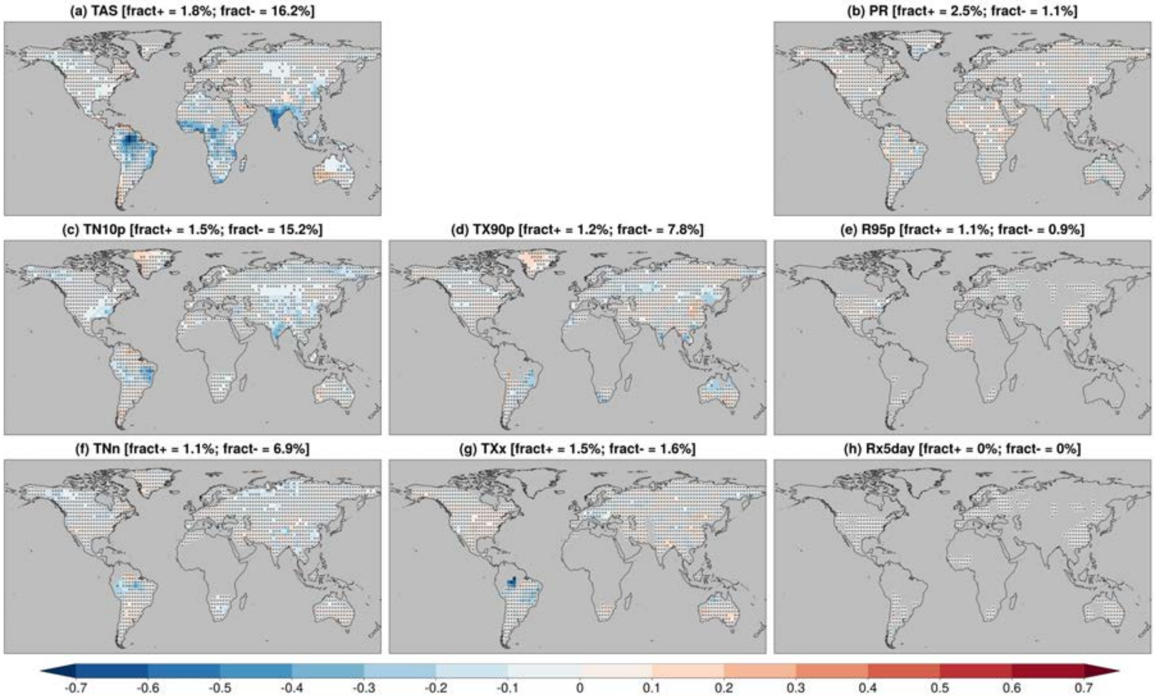

(a) TAS [fract+ = 1.8%; fract- = 16.2%]
(b) PR [fract+ = 2.5%; fract- = 1.1%]
(c) TN10p [fract+ = 1.5%; fract- = 15.2%]
(d) TX90p [fract+ = 1.2%; fract- = 7.8%]
(e) R95p [fract+ = 1.1%; fract- = 0.9%]
(f) TNn [fract+ = 1.1%; fract- = 6.9%]
(g) TXx [fract+ = 1.5%; fract- = 1.6%]
(h) Rx5day [fract+ = 0%; fract- = 0%]
-0.7
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

Figure S6. As Figure S4, but using HadEX3 as the reference dataset for the extreme indices.

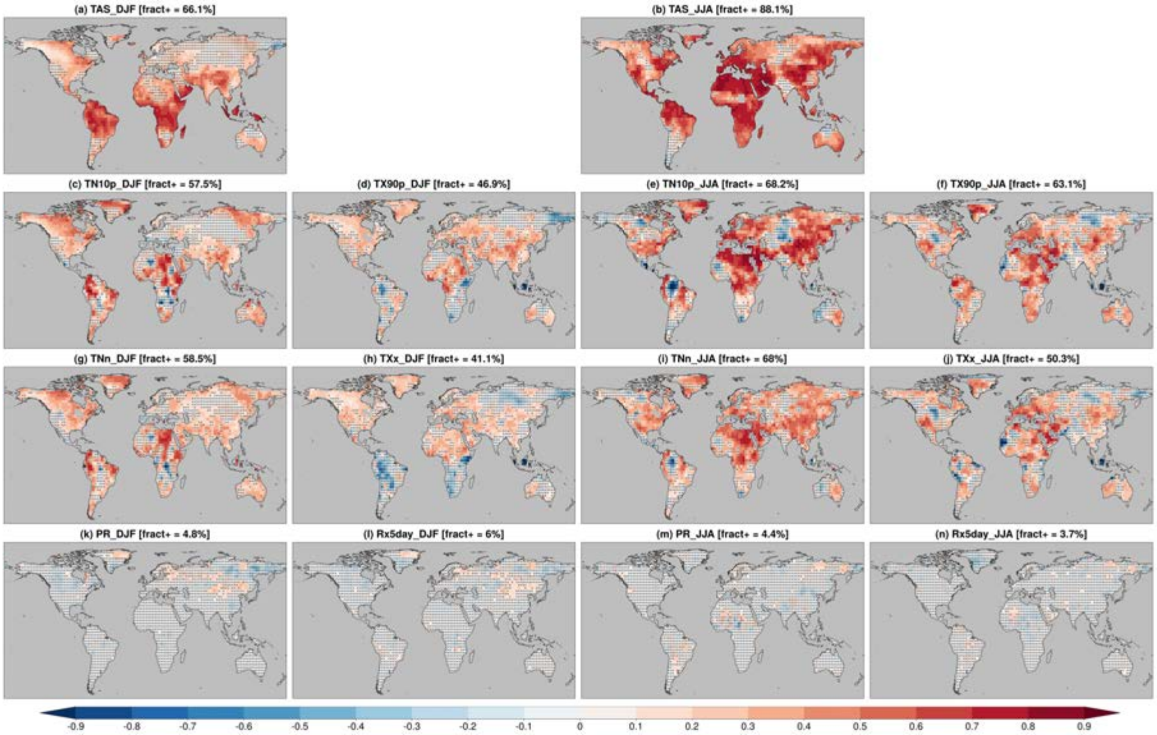

(a) TAS_DJF [fract+ = 66.1%]
(b) TAS_JJA [fract+ = 88.1%]
(c) TN10p_DJF [fract+ = 57.5%]
(d) TX90p_DJF [fract+ = 46.9%]
(e) TN10p_JJA [fract+ = 68.2%]
(f) TX90p_JJA [fract+ = 63.1%]
(g) TNn_DJF [fract+ = 58.5%]
(h) TXx_DJF [fract+ = 41.1%]
(i) TNn_JJA [fract+ = 68%]
(j) TXx_JJA [fract+ = 50.3%]
(k) PR_DJF [fract+ = 4.8%]
(l) Rx5day_DJF [fract+ = 6%]
(m) PR_JJA [fract+ = 4.4%]
(n) Rx5day_JJA [fract+ = 3.7%]
-0.9
-0.8
-0.7
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

Figure S7. As Figure 3, but for the RPSS of the DCPP multi-model ensemble using the climatological forecast as the reference forecast. The statistical significance has been estimated with the Random Walk test controlling the FDR with $\alpha_{FDR} = 0.1$.

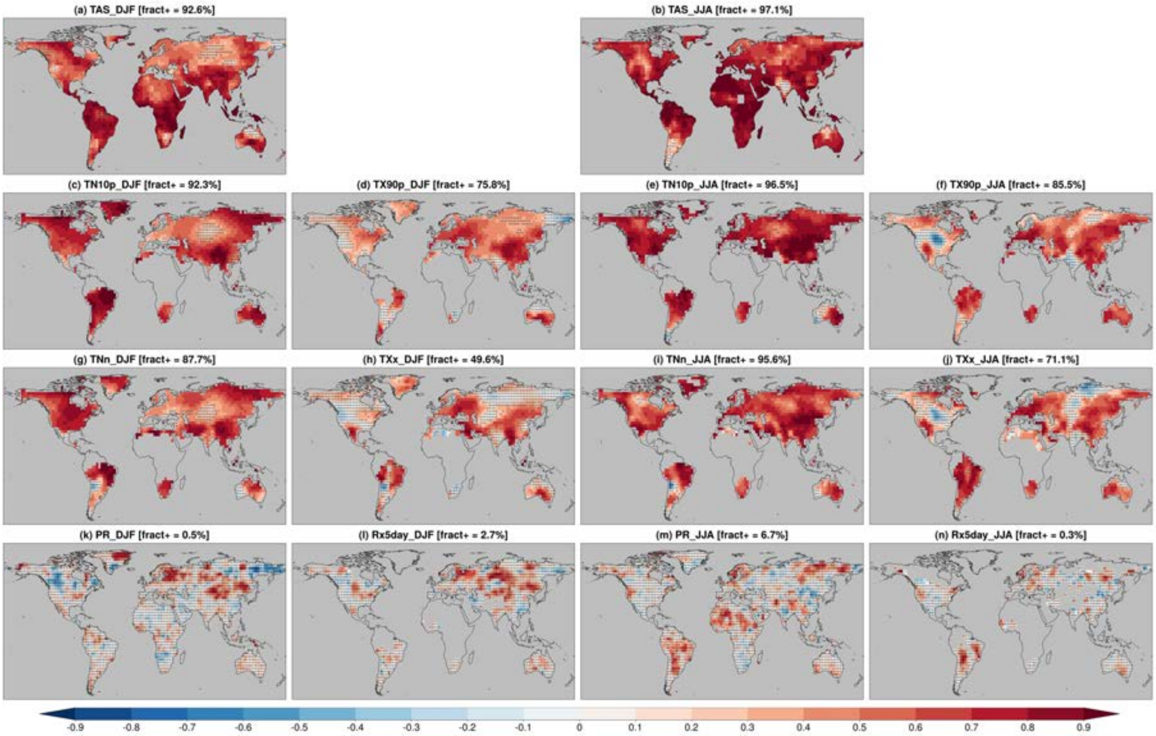

(a) TAS_DJF [fract+ = 92.6%]
(b) TAS_JJA [fract+ = 97.1%]
(c) TN10p_DJF [fract+ = 92.3%]
(d) TX90p_DJF [fract+ = 75.8%]
(e) TN10p_JJA [fract+ = 96.5%]
(f) TX90p_JJA [fract+ = 85.5%]
(g) TNn_DJF [fract+ = 87.7%]
(h) TXx_DJF [fract+ = 49.6%]
(i) TNn_JJA [fract+ = 95.6%]
(j) TXx_JJA [fract+ = 71.1%]
(k) PR_DJF [fract+ = 0.5%]
(l) Rx5day_DJF [fract+ = 2.7%]
(m) PR_JJA [fract+ = 6.7%]
(n) Rx5day_JJA [fract+ = 0.3%]
-0.9 -0.8 -0.7 -0.6 -0.5 -0.4 -0.3 -0.2 -0.1 0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9

Figure S8. As Figure 3, but using the HadEX3 as the reference dataset for the extreme indices.

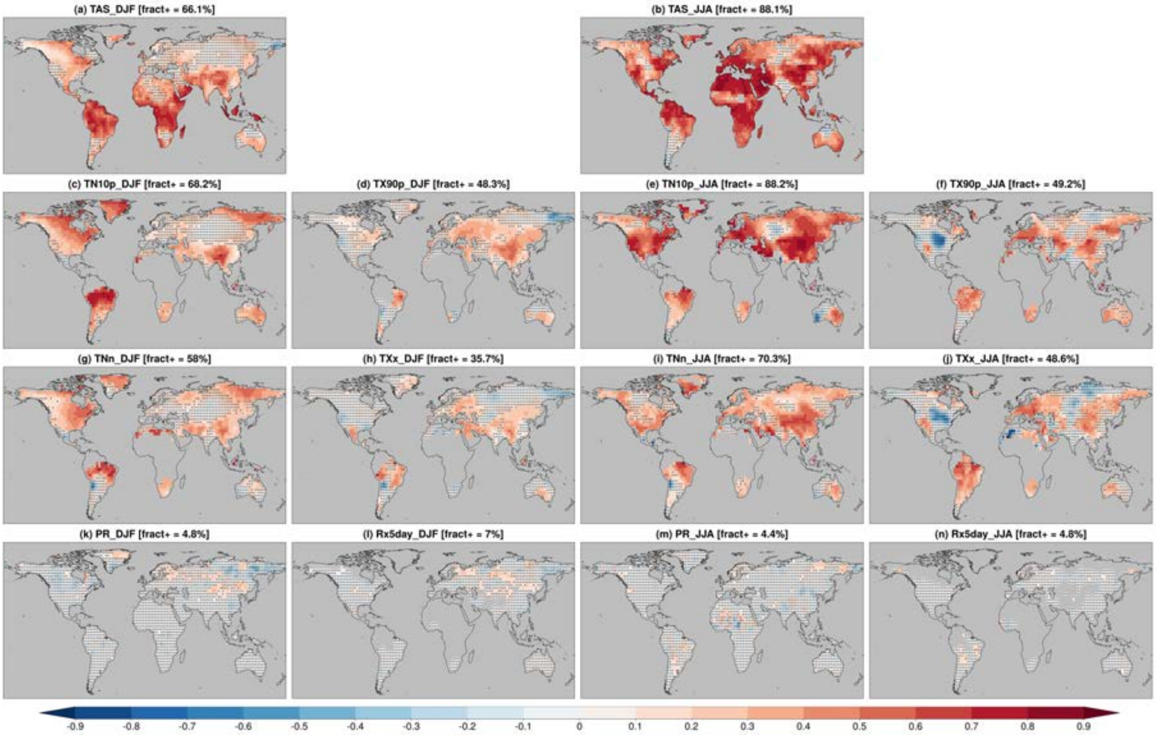

(a) TAS_DJF [fract+ = 66.1%]
(b) TAS_JJA [fract+ = 88.1%]
(c) TN10p_DJF [fract+ = 68.2%]
(d) TX90p_DJF [fract+ = 48.3%]
(e) TN10p_JJA [fract+ = 88.2%]
(f) TX90p_JJA [fract+ = 49.2%]
(g) TNn_DJF [fract+ = 58%]
(h) TXx_DJF [fract+ = 35.7%]
(i) TNn_JJA [fract+ = 70.3%]
(j) TXx_JJA [fract+ = 48.6%]
(k) PR_DJF [fract+ = 4.8%]
(l) Rx5day_DJF [fract+ = 7%]
(m) PR_JJA [fract+ = 4.4%]
(n) Rx5day_JJA [fract+ = 4.8%]
-0.9 -0.8 -0.7 -0.6 -0.5 -0.4 -0.3 -0.2 -0.1 0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9

Figure S9. As Figure S7, but using HadEX3 as the reference dataset for the extreme indices.

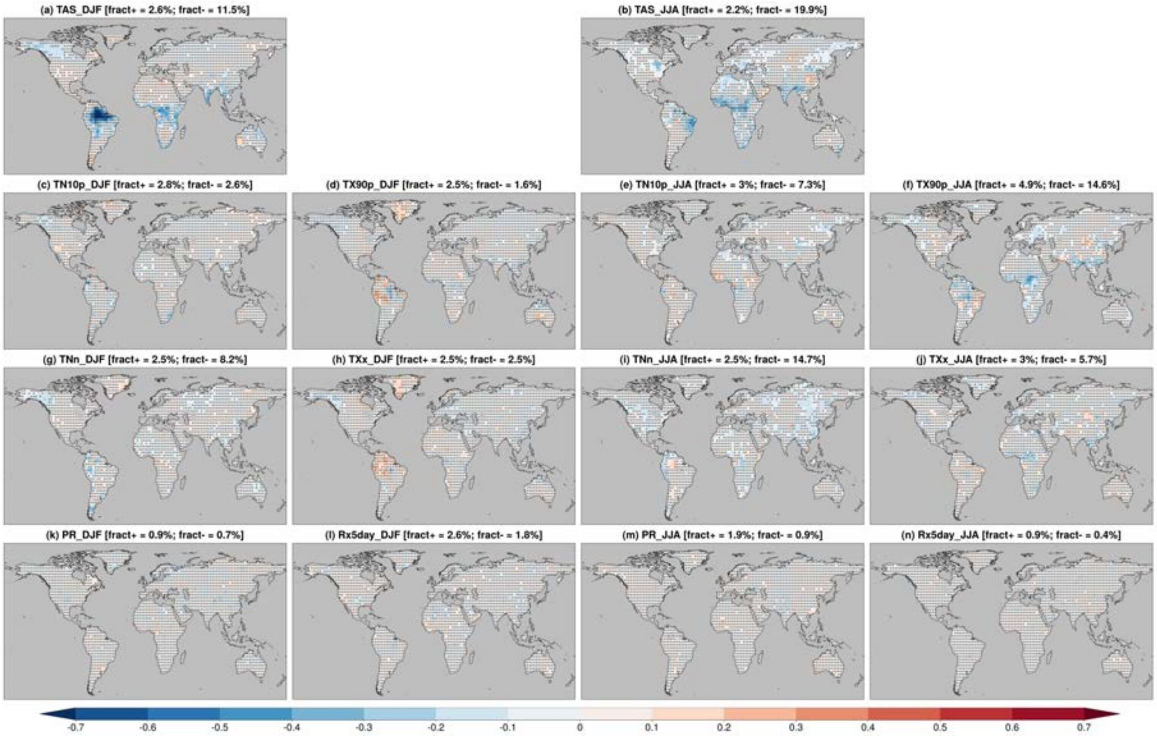

(a) TAS_DJF [fract+ = 2.6%; fract- = 11.5%]
(b) TAS_JJA [fract+ = 2.2%; fract- = 19.9%]
(c) TN10p_DJF [fract+ = 2.8%; fract- = 2.6%]
(d) TX90p_DJF [fract+ = 2.5%; fract- = 1.6%]
(e) TN10p_JJA [fract+ = 3%; fract- = 7.3%]
(f) TX90p_JJA [fract+ = 4.9%; fract- = 14.6%]
(g) TNn_DJF [fract+ = 2.5%; fract- = 8.2%]
(h) TXx_DJF [fract+ = 2.5%; fract- = 2.5%]
(i) TNn_JJA [fract+ = 2.5%; fract- = 14.7%]
(j) TXx_JJA [fract+ = 3%; fract- = 5.7%]
(k) PR_DJF [fract+ = 0.9%; fract- = 0.7%]
(l) Rx5day_DJF [fract+ = 2.6%; fract- = 1.8%]
(m) PR_JJA [fract+ = 1.9%; fract- = 0.9%]
(n) Rx5day_JJA [fract+ = 0.9%; fract- = 0.4%]
-0.7
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

Figure S10. As Figure 4, but for the RPSS of the DCPP multi-model ensemble using the HIST multi-model ensemble as the reference forecast. The statistical significance has been estimated with the Random Walk test controlling the FDR with $\alpha_{FDR} = 0.1$.

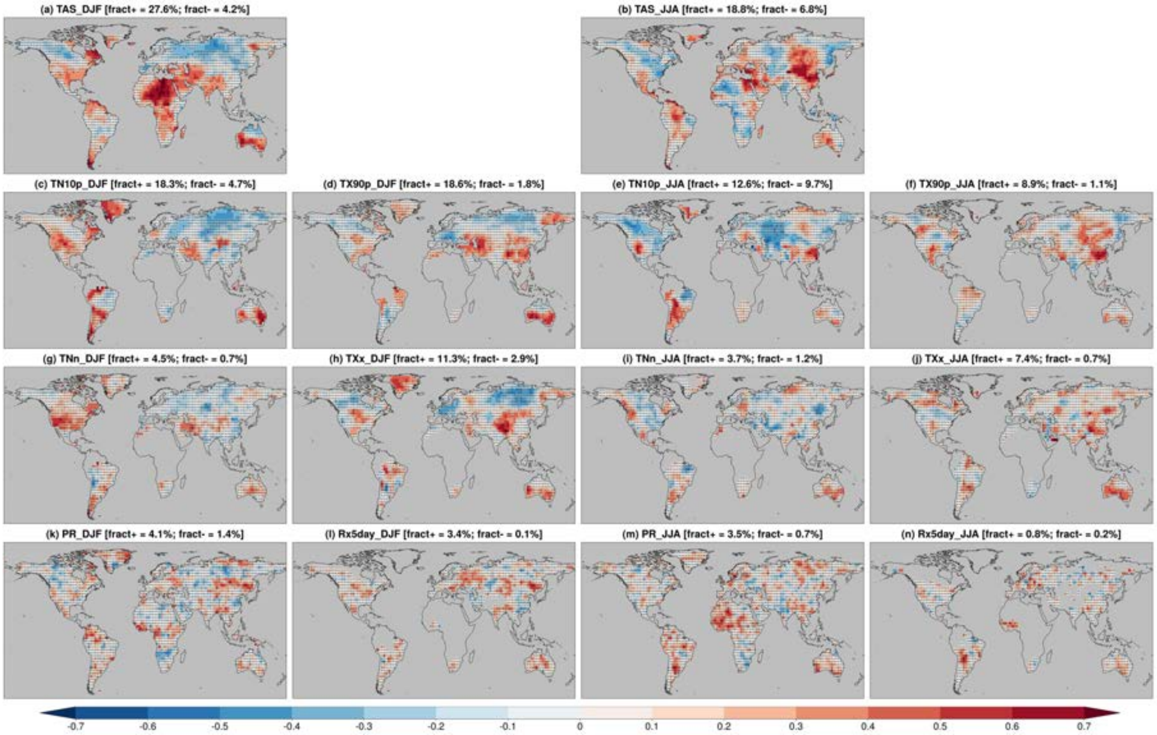

(a) TAS_DJF [fract+ = 27.6%; fract- = 4.2%]
(b) TAS_JJA [fract+ = 18.8%; fract- = 6.8%]
(c) TN10p_DJF [fract+ = 18.3%; fract- = 4.7%]
(d) TX90p_DJF [fract+ = 18.6%; fract- = 1.8%]
(e) TN10p_JJA [fract+ = 12.6%; fract- = 9.7%]
(f) TX90p_JJA [fract+ = 8.9%; fract- = 1.1%]
(g) TNn_DJF [fract+ = 4.5%; fract- = 0.7%]
(h) TXx_DJF [fract+ = 11.3%; fract- = 2.9%]
(i) TNn_JJA [fract+ = 3.7%; fract- = 1.2%]
(j) TXx_JJA [fract+ = 7.4%; fract- = 0.7%]
(k) PR_DJF [fract+ = 4.1%; fract- = 1.4%]
(l) Rx5day_DJF [fract+ = 3.4%; fract- = 0.1%]
(m) PR_JJA [fract+ = 3.5%; fract- = 0.7%]
(n) Rx5day_JJA [fract+ = 0.8%; fract- = 0.2%]
-0.7
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

Figure S11. As Figure 4, but using HadEX3 as the reference dataset for the extreme indices.

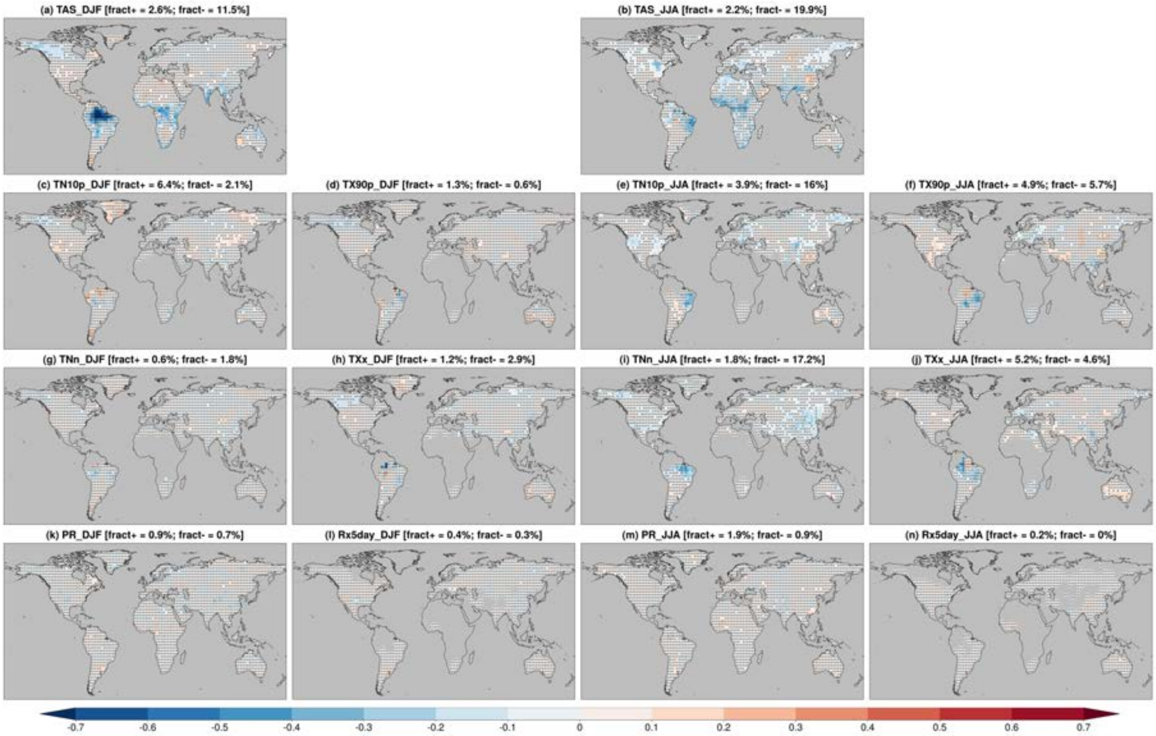

(a) TAS_DJF [fract+ = 2.6%; fract- = 11.5%]
(b) TAS_JJA [fract+ = 2.2%; fract- = 19.9%]
(c) TN10p_DJF [fract+ = 6.4%; fract- = 2.1%]
(d) TX90p_DJF [fract+ = 1.3%; fract- = 0.6%]
(e) TN10p_JJA [fract+ = 3.9%; fract- = 16%]
(f) TX90p_JJA [fract+ = 4.9%; fract- = 5.7%]
(g) TNn_DJF [fract+ = 0.6%; fract- = 1.8%]
(h) TXx_DJF [fract+ = 1.2%; fract- = 2.9%]
(i) TNn_JJA [fract+ = 1.8%; fract- = 17.2%]
(j) TXx_JJA [fract+ = 5.2%; fract- = 4.6%]
(k) PR_DJF [fract+ = 0.9%; fract- = 0.7%]
(l) Rx5day_DJF [fract+ = 0.4%; fract- = 0.3%]
(m) PR_JJA [fract+ = 1.9%; fract- = 0.9%]
(n) Rx5day_JJA [fract+ = 0.2%; fract- = 0%]
-0.7
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

Figure S12. As Figure S10, but using HadEX3 as the reference dataset for the extreme indices.

References

Bethke I, Wang Y, Counillon F, Keenlyside N, Kimmritz M, Fransner F, Samuelsen A, Langehaug H, Svendsen L, Chiu P G, Passos L, Bentsen M, Guo C, Gupta A, Tjiputra J, Kirkevåg A, Olivié D, Øyvind Seland, Vågane J S, Fan Y & Eldevik T 2021 *Geoscientific Model Development* **14**, 7073–7116.

Boucher O, Servonnat J, Albright A L, Aumont O, Balkanski Y, Bastrikov V, Bekki S, Bonnet R, Bony S, Bopp L, Braconnot P, Brockmann P, Cadule P, Caubel A, Cheruy F, Codron F, Cozic A, Cugnet D, D'Andrea F, Davini P, de Lavergne C, Denvil S, Deshayes J, Devilliers M, Ducharne A, Dufresne J L, Dupont E, Éthé C, Fairhead L, Falletti L, Flavoni S, Foujols M A, Gardoll S, Gastineau G, Ghattas J, Grandpeix J Y, Guenet B, Lionel E G, Guilyardi E, Guimberteau M, Hauglustaine D, Hourdin F, Idelkadi A, Joussaume S, Kageyama M, Khodri M, Krinner G, Lebas N, Levavasseur G, Lévy C, Li L, Lott F, Lurton T, Luyssaert S, Madec G, Madeleine J B, Maignan F, Marchand M, Marti O, Mellul L, Meurdesoif Y, Mignot J, Musat I, Ottlé C, Peylin P, Planton Y, Polcher J, Rio C, Rochetin N, Rousset C, Sepulchre P, Sima A, Swingedouw D, Thiéblemont R, Traore A K, Vancoppenolle M, Vial J, Vialard J, Viovy N & Vuichard N 2020 *Journal of Advances in Modeling Earth Systems* **12**, e2019MS002010.

Contractor S, Donat M G, Alexander L V, Ziese M, Meyer-Christoffer A, Schneider U, Rustemeier E, Becker A, Durre I & Vose R S 2020 *Hydrology and Earth System Sciences* **24**, 919–943.

Dunn R J, Alexander L V, Donat M G, Zhang X, Bador M, Herold N, Lippmann T, Allan R, Aguilar E, Barry A A, Brunet M, Caesar J, Chagnaud G, Cheng V, Cinco T, Durre I, de Guzman R, Htay T M, Ibadullah W M W, Ibrahim M K I B, Khoshkam M, Kruger A, Kubota H, Leng T W, Lim G, Li-Sha L, Marengo J, Mbatha S, McGree S, Menne M, de los Milagros Skansi M, Ngwenya S, Nkrumah F, Oonariya C, Pabon-Caicedo J D, Panthou G, Pham C, Rahimzadeh F, Ramos A, Salgado E, Salinger J, Sané Y, Sopaheluwakan A, Srivastava A, Sun Y, Timbal B, Trachow N, Trewin B, van der Schrier G, Vazquez-Aguirre J, Vasquez R, Villarroel C, Vincent L, Vischel T, Vose R & Yussof M N A B H 2020 *Journal of Geophysical Research: Atmospheres* **125**, e2019JD032263.

Döscher R, Acosta M, Alessandri A, Anthoni P, Arsouze T, Bergman T, Bernardello R, Boussetta S, Caron L P, Carver G, Castrillo M, Catalano F, Cvijanovic I, Davini P, Dekker E, Doblas-Reyes F J, Docquier D, Echevarria P, Fladrich U, Fuentes-Franco R, Gröger M, Hardenberg J V, Hieronymus J, Karami M P, Keskinen J P, Koenigk T, Makkonen R, Massonnet F, Ménégoz M, Miller P A, Moreno-Chamarro E, Nieradzik L, Noije T V, Nolan P, O'donnell D, Ollinaho P, Oord G V D, Ortega P, Prims O T, Ramos A, Reerink T, Rousset C, Ruprich-Robert Y, Sager P L, Schmith T, Schrödner R, Serva F, Sicardi V, Madsen M S, Smith B, Tian T, Tourigny E, Uotila P, Vancoppenolle M, Wang S, Wårlind D, Willén U, Wyser K, Yang S, Yepes-Arbós X & Zhang Q 2022 *Geoscientific Model Development* **15**, 2973–3020.

Menne M J, Williams C N, Gleason B E, Rennie J J & Lawrimore J H 2018 *Journal of Climate* **31**, 9835–9854.

Müller W A, Jungclaus J H, Mauritsen T, Baehr J, Bittner M, Budich R, Bunzel F, Esch M, Ghosh R, Haak H, Ilyina T, Kleine T, Kornblueh L, Li H, Modali K, Notz D, Pohlmann H, Roeckner E, Stemmler I, Tian F & Marotzke J 2018 *Journal of Advances in Modeling Earth Systems* **10**, 1383–1413.

Nicolì D, Bellucci A, Ruggieri P, Athanasiadis P J, Materia S, Peano D, Fedele G, Hénin R & Gualdi S 2023 *Geoscientific Model Development* **16**, 179–197.

Schneider U, Becker A, Finger P, Rustemeier E & Ziese M 2020 *Global Precipitation Climatology Centre (GPCC, http://gpcc.dwd.de/) at Deutscher Wetterdienst* .

Sellar A A, Walton J, Jones C G, Wood R, Abraham N L, Andrejczuk M, Andrews M B, Andrews T, Archibald A T, de Mora L, Dyson H, Elkington M, Ellis R, Florek P, Good P, Gohar L, Haddad S, Hardiman S C, Hogan E, Iwi A, Jones C D, Johnson B, Kelley D I, Kettleborough J, Knight J R, Köhler M O, Kuhlbrodt T, Liddicoat S, Linova-Pavlova I, Mizielinski M S, Morgenstern O, Mulcahy J, Neininger E, O'Connor F M, Petrie R, Ridley J, Rioual J C, Roberts M, Robertson E, Rumbold S, Seddon J, Shepherd H, Shim S, Stephens A, Teixiera J C, Tang Y, Williams J, Wiltshire A & Griffiths P T 2020 *Journal of Advances in Modeling Earth Systems* **12**, e2019MS001946.

Swart N C, Cole J N, Kharin V V, Lazare M, Scinocca J F, Gillett N P, Anstey J, Arora V, Christian J R, Hanna S, Jiao Y, Lee W G, Majaess F, Saenko O A, Seiler C, Seinen C, Shao A, Sigmond M, Solheim L, Salzen K V, Yang D & Winter B 2019 *Geoscientific Model Development* **12**, 4823–4873.

Tatebe H, Ogura T, Nitta T, Komuro Y, Ogochi K, Takemura T, Sudo K, Sekiguchi M, Abe M, Saito F, Chikira M, Watanabe S, Mori M, Hirota N, Kawatani Y, Mochizuki T, Yoshimura K, Takata K, O'Ishi R, Yamazaki D, Suzuki T, Kurogi M, Kataoka T, Watanabe M & Kimoto M 2019 *Geoscientific Model Development* **12**, 2727–2765.

Wu T, Lu Y, Fang Y, Xin X, Li L, Li W, Jie W, Zhang J, Liu Y, Zhang L, Zhang F, Zhang Y, Wu F, Li J, Chu M, Wang Z, Shi X, Liu X, Wei M, Huang A, Zhang Y & Liu X 2019 *Geoscientific Model Development* **12**, 1573–1600.

Yukimoto S, Kawai H, Koshiro T, Oshima N, Yoshida K, Urakawa S, Tsujino H, Deushi M, Tanaka T, Hosaka M, Yabu S, Yoshimura H, Shindo E, Mizuta R, Obata A, Adachi Y & Ishii M 2019 *Journal of the Meteorological Society of Japan. Ser. II* **97**, 931–965.

References

Abbass, K., Qasim, M.Z., Song, H., Murshed, M., Mahmood, H. & Younis, I. (2022). A review of the global climate change impacts, adaptation, and sustainable mitigation measures. *Environmental Science and Pollution Research*, **29**, 42539–42559. 1

Adler, R.F., Huffman, G.J., Chang, A., Ferraro, R., Xie, P.P., Janowiak, J., Rudolf, B., Schneider, U., Curtis, S., Bolvin, D., Gruber, A., Susskind, J., Arkin, P. & Nelkin, E. (2003). The version-2 global precipitation climatology project (gpcp) monthly precipitation analysis (1979–present). *Journal of Hydrometeorology*, **4**, 1147 – 1167. 13

Ashok, K., Guan, Z. & Yamagata, T. (2001). Impact of the indian ocean dipole on the relationship between the indian monsoon rainfall and enso. *Geophysical Research Letters*, **28**, 4499–4502. 5

Athanasiadis, P.J., Yeager, S., Kwon, Y.O., Bellucci, A., Smith, D.W. & Tibaldi, S. (2020). Decadal predictability of north atlantic blocking and the nao. *npj Climate and Atmospheric Science 2020 3:1*, **3**, 1–10. 11

Bauer, P., Thorpe, A. & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55. 6

Baulenas, E., Bojovic, D., Urquiza, D., Terrado, M., Pickard, S., González, N. & Clair, A.L.S. (2023). User selection and engagement for climate services coproduction. *Weather, Climate, and Society*, **15**, 381 – 392. 15

Bellucci, A., Haarsma, R., Gualdi, S., Athanasiadis, P.J., Caian, M., Cassou, C., Fernandez, E., Germe, A., Jungclaus, J., Kröger, J., Matei, D., Müller, W., Pohlmann, H., y Melia, D.S., Sanchez, E., Smith, D., Terray, L., Wyser, K. & Yang, S. (2014). An assessment of a multi-model ensemble of decadal climate predictions. *Climate Dynamics 2014 44:9*, **44**, 2787–2806. 14

Benestad, R., Caron, L.P., Parding, K., Iturbide, M., Llorente, J.M.G., Mezghani, A. & Doblas-Reyes, F.J. (2019). Using statistical downscaling to assess skill of decadal predictions. *Tellus A: Dynamic Meteorology and Oceanography*. 14

Bilbao, R., Wild, S., Ortega, P., Acosta-Navarro, J., Arsouze, T., Bretonnière, P.A., Caron, L.P., Castrillo, M., Cruz-García, R., Cvijanovic, I., Doblas-Reyes, F.J., Donat, M., Dutra, E., Echevarría, P., Ho, A.C., Loosveldt-Tomas, S., Moreno-Chamarro, E., Pérez-Zanon, N., Ramos, A., Ruprich-Robert, Y., Sicardi, V., Tourigny, E. & Vegas-Regidor, J. (2021). Assessment of a full-field initialized decadal climate prediction system with the cmip6 version of ec-earth. *Earth System Dynamics*, **12**, 173–196. 10, 18

Boer, G.J., Smith, D.M., Cassou, C., Doblas-Reyes, F., Danabasoglu, G., Kirtman, B., Kushnir, Y., Kimoto, M., Meehl, G.A., Msadek, R., Mueller, W.A., Taylor, K.E., Zwiers, F., Rixen, M., Ruprich-Robert, Y. & Eade, R. (2016). The decadal climate prediction project (dcpp) contribution to cmip6. *Geoscientific Model Development*, **9**, 3751–3777. 16

Bojovic, D., St. Clair, A.L., Christel, I., Terrado, M., Stanzel, P., Gonzalez, P. & Palin, E.J. (2021). Engagement, involvement and empowerment: Three realms of a coproduction framework for climate services. *Global Environmental Change*, **68**, 102271. 15, 16

Brovkin, V., Sitch, S., Von Bloh, W., Claussen, M., Bauer, E. & Cramer, W. (2004). Role of land cover changes for atmospheric co2 increase and climate change during the last 150 years. *Global Change Biology*, **10**, 1253–1266. 6

BrunoSoares, M., Alexander, M. & Dessai, S. (2018). Sectoral use of climate information in europe: A synoptic overview. *Climate Services*, **9**, 5–20. 8, 15, 16

Caron, L.P., Hermanson, L., Dobbin, A., Imbers, J., Lledó, L. & Vecchi, G.A. (2018). How skillful are the multiannual forecasts of atlantic hurricane activity? *Bulletin of the American Meteorological Society*, **99**, 403–413. 15

Cassou, C., Kushnir, Y., Hawkins, E., Pirani, A., Kucharski, F., Kang, I.S. & Caltabiano, N. (2018). Decadal climate variability and predictability: Challenges and opportunities. *Bulletin of the American Meteorological Society*, **99**, 479 – 490. 10

Chen, M., Wang, W. & Kumar, A. (2013). Lagged ensembles, forecast configuration, and seasonal predictions. *Monthly Weather Review*, **141**, 3477 – 3497. 9

Chiang, Y.C. & Ling, T.Y. (2017). Exploring flood resilience thinking in the retail sector under climate change: a case study of an estuarine region of taipei city. *Sustainability*, **9**, 1650. 15

Curtis, S., Fair, A., Wistow, J., Val, D.V. & Oven, K. (2017). Impact of extreme weather events and climate change for health and social care systems. *Environmental Health: A Global Access Science Source*, **16**, 23–32. 1

Delgado-Torres, C., Donat, M.G., Gonzalez-Reviriego, N., Caron, L.P., Athanasiadis, P.J., Bretonnière, P.A., Dunstone, N.J., Ho, A.C., Nicoli, D., Pankatz, K., Paxian, A., Pérez-Zanón, N., Cabré, M.S., Solaraju-Murali, B., Soret, A. & Doblas-Reyes, F.J. (2022a). Multi-model forecast quality assessment of cmip6 decadal predictions. *Journal of Climate*, **35**, 4363 – 4382. 73, 76

Delgado-Torres, C., Verfaillie, D., Mohino, E. & Donat, M.G. (2022b). Representation and annual to decadal predictability of euro-atlantic weather regimes in the cmip6 version of the ec-earth coupled climate model. *Journal of Geophysical Research: Atmospheres*, **127**, e2022JD036673. 73

Delgado-Torres, C., Donat, M.G., Soret, A., González-Reviriego, N., Bretonnière, P.A., Ho, A.C., Pérez-Zanón, N., Cabré, M.S. & Doblas-Reyes, F.J. (2023). Multi-annual predictions of the frequency and intensity of daily temperature and precipitation extremes. *Environmental Research Letters*, **18**, 034031. 73

DelSole, T., Nattala, J. & Tippett, M.K. (2014). Skill improvement from increased ensemble size and model diversity. *Geophysical Research Letters*, **41**, 7331–7342. 14

Deser, C., Phillips, A., Bourdette, V. & Teng, H. (2012). Uncertainty in climate change projections: the role of internal variability. *Climate dynamics*, **38**, 527–546. 2

Doblas-Reyes, F.J., Hagedorn, R. & Palmer, T. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting — ii. calibration and combination. *http://dx.doi.org/10.3402/tellusa.v57i3.14658*, **57**, 234–252. 14

Doblas-Reyes, F.J., Andreu-Burillo, I., Chikamoto, Y., García-Serrano, J., Guemas, V., Kimoto, M., Mochizuki, T., Rodrigues, L.R.L. & van Oldenborgh, G.J. (2013). Initialized near-term regional climate change prediction. *Nature Communications 2013 4:1*, **4**, 1–9. 5, 7

Donat, M.G., Delgado-Torres, C., De Luca, P., Mahmood, R., Ortega, P. & Doblas-Reyes, F.J. (2023). How credibly do cmip6 simulations capture historical mean and extreme precipitation changes? *Geophysical Research Letters*, **50**, e2022GL102466, e2022GL102466 2022GL102466. 11

Dunstone, N., Smith, D., Yeager, S., Danabasoglu, G., Monerie, P.A., Hermanson, L., Eade, R., Ineson, S., Robson, J., Scaife, A. & Ren, H.L. (2020). Skilful interannual climate prediction from two large initialised model ensembles. *Environmental Research Letters*, **15**, 094083. 8

Eade, R., Smith, D., Scaife, A., Wallace, E., Dunstone, N., Hermanson, L. & Robinson, N. (2014). Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophysical Research Letters*, **41**, 5620–5628. 10, 14

Eyring, V., Bony, S., Meehl, G.A., Senior, C.A., Stevens, B., Stouffer, R.J. & Taylor, K.E. (2016). Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, **9**, 1937–1958. 16

Frumkin, H., Hess, J., Luber, G., Malilay, J. & McGeehin, M. (2011). Climate change: The public health response. *https://doi.org/10.2105/AJPH.2007.119362*, **98**, 435–445. 15

Gangstø, R., Weigel, A.P., Liniger, M.A. & Appenzeller, C. (2013). Methodological aspects of the validation of decadal predictions. *Climate Research*, **55**, 181–200. 13, 14

Goddard, L., Hurrell, J.W., Kirtman, B.P., Murphy, J., Stockdale, T. & Vera, C. (2012). Two time scales for the price of one (almost). *Bulletin of the American Meteorological Society*, **93**, 621 – 629. 12, 15

Goddard, L., Kumar, A., Solomon, A., Smith, D., Boer, G., Gonzalez, P., Kharin, V., Merryfield, W., Deser, C., Mason, S.J., Kirtman, B.P., Msadek, R., Sutton, R., Hawkins, E., Fricker, T., Hegerl, G., Ferro, C.A., Stephenson, D.B., Meehl, G.A., Stockdale, T., Burgman, R., Greene, A.M., Kushnir, Y., Newman, M., Carton, J., Fukumori, I. & Delworth, T. (2013). A verification framework for interannual-to-decadal predictions experiments. *Climate Dynamics*, **40**. 12, 13

Hagedorn, R., Doblas-Reyes, F.J. & Palmer, T. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting — i. basic concept. *http://dx.doi.org/10.3402/tellusa.v57i3.14657*, **57**, 219–233. 14

Hanlon, H.M., Hegerl, G.C., Tett, S.F. & Smith, D.M. (2013). Can a decadal forecasting system predict temperature extreme indices? *Journal of Climate*, **26**, 3728–3744. 1

HAWKINS, E. & SUTTON, R. (2009). The potential to narrow uncertainty in regional climate predictions. *Bulletin of the American Meteorological Society*, **90**, 1095 – 1108. 10

HAZELEGER, W., GUEMAS, V., WOUTERS, B., CORTI, S., ANDREU–BURILLO, I., DOBLAS–REYES, F.J., WYSER, K. & CAIAN, M. (2013). Multiyear climate predictions using two initialization strategies. *Geophysical Research Letters*, **40**, 1794–1798. 9

HEMRI, S., BHEND, J., LINIGER, M.A., MANZANAS, R., SIEGERT, S., STEPHENSON, D.B., GUTIÉRREZ, J.M., BROOKSHAW, A. & DOBLAS-REYES, F.J. (2020). How to create an operational multi-model of seasonal forecasts? *Climate Dynamics 2020 55:5*, **55**, 1141–1157. 14

HENLEY, B.J., GERGIS, J., KAROLY, D.J., POWER, S., KENNEDY, J. & FOLLAND, C.K. (2015). A tripole index for the interdecadal pacific oscillation. *Climate dynamics*, **45**, 3077–3090. 5

HERMANSON, L., BILBAO, R., DUNSTONE, N., MÉNÉGOZ, M., ORTEGA, P., POHLMANN, H., ROBSON, J.I., SMITH, D.M., STRAND, G., TIMMRECK, C., YEAGER, S. & DANABASOGLU, G. (2020). Robust multiyear climate impacts of volcanic eruptions in decadal prediction systems. *Journal of Geophysical Research: Atmospheres*, **125**, e2019JD031739, e2019JD031739 10.1029/2019JD031739. 3, 6

HERSBACH, H., BELL, B., BERRISFORD, P., HIRAHARA, S., HORÁNYI, A., MUÑOZ-SABATER, J., NICOLAS, J., PEUBEY, C., RADU, R., SCHEPERS, D., SIMMONS, A., SOCI, C., ABDALLA, S., ABELLAN, X., BALSAMO, G., BECHTOLD, P., BIAVATI, G., BIDLOT, J., BONAVITA, M., DE CHIARA, G., DAHLGREN, P., DEE, D., DIAMANTAKIS, M., DRAGANI, R., FLEMMING, J., FORBES, R., FUENTES, M., GEER, A., HAIMBERGER, L., HEALY, S., HOGAN, R.J., HÓLM, E., JANISKOVÁ, M., KEELEY, S., LALOYAUX, P., LOPEZ, P., LUPU, C., RADNOTI, G., DE ROSNAY, P., ROZUM, I., VAMBORG, F., VILLAUME, S. & THÉPAUT, J.N. (2020). The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, **146**, 1999–2049. 13

Hopson, T.M. (2014). Assessing the ensemble spread-error relationship. *Monthly Weather Review*, **142**, 1125–1142. 74

IPCC (2023a). Future global climate: Scenario-based projections and near-term information. *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, 553–672. 7

IPCC (2023b). Summary for policymakers. *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, 1513–1766. 1

Jain, P. (1993). Greenhouse effect and climate change: scientific basis and overview. *Renewable Energy*, **3**, 403–420, solar radiation, environment and climate change. 6

Jia, L., Yang, X., Vecchi, G.A., Gudgel, R.G., Delworth, T.L., Rosati, A., Stern, W.F., Wittenberg, A.T., Krishnamurthy, L., Zhang, S., Msadek, R., Kapnick, S., Underwood, S., Zeng, F., Anderson, W.G., Balaji, V. & Dixon, K. (2015). Improved seasonal prediction of temperature and precipitation over land in a high-resolution gfdl climate model. *Journal of Climate*, **28**, 2044 – 2062. 10

Jolliffe, I.T. & Stephenson, D.B. (2012). *Forecast Verification: A Practitioner's Guide in Atmospheric Science, Second Edition*. John Wiley I& Sons, Ltd. 13

Kenyon, J. & Hegerl, G.C. (2008). Influence of modes of climate variability on global temperature extremes. *Journal of Climate*, **21**, 3872 – 3889. 3

Kharin, V.V. & Zwiers, F.W. (2003). On the roc score of probability forecasts. *Journal of Climate*, **16**, 4145 – 4150. 74

Kidson, J.W. (1988). Interannual variations in the southern hemisphere circulation. *Journal of Climate*, **1**, 1177 – 1198. 4

Kim, Y.H. & Chun, H.Y. (2015). Momentum forcing of the quasi-biennial oscillation by equatorial waves in recent reanalyses. *Atmospheric Chemistry and Physics*, **15**, 6577–6587. 5

Kirtman, B., Power, S.B. & et al. (2013). Near-term climate change: Projections and predictability. *In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. 7, 8, 9

Kröger, J., Pohlmann, H., Sienz, F., Marotzke, J., Baehr, J., Köhl, A., Modali, K., Polkova, I., Stammer, D., Vamborg, F.S.E. *et al.* (2018). Full-field initialized decadal predictions with the mpi earth system model: An initial shock in the north atlantic. *Climate dynamics*, **51**, 2593–2608. 10

Kushnir, Y., Scaife, A.A., Arritt, R., Balsamo, G., Boer, G., Doblas-Reyes, F., Hawkins, E., Kimoto, M., Kolli, R.K., Kumar, A., Matei, D., Matthes, K., Müller, W.A., O'Kane, T., Perlwitz, J., Power, S., Raphael, M., Shimpo, A., Smith, D., Tuma, M. & Wu, B. (2019). Towards operational predictions of the near-term climate. *Nature Climate Change 2019 9:2*, **9**, 94–101. 1, 7

L'Heureux, M.L., Tippett, M.K., Kumar, A., Butler, A.H., Ciasto, L.M., Ding, Q., Harnos, K.J. & Johnson, N.C. (2017). Strong relations between enso and the arctic oscillation in the north american multimodel ensemble. *Geophysical Research Letters*, **44**, 11,654–11,662. 3

Lorenz, E.N. (1963). Deterministic nonperiodic flow. *Journal of Atmospheric Sciences*, **20**, 130 – 141. 9

Mantua, N.J. & Hare, S.R. (2002). The pacific decadal oscillation. *Journal of oceanography*, **58**, 35–44. 5

Manzanas, R., Gutiérrez, J.M., Bhend, J., Hemri, S., Doblas-Reyes, F.J., Torralba, V., Penabad, E. & Brookshaw, A. (2019). Bias adjustment and ensemble recalibration methods for seasonal forecasting: a comprehensive intercomparison using the c3s dataset. *Climate Dynamics 2019 53:3*, **53**, 1287–1305. 14

Marcos, R., Llasat, M.C., Quintana-Seguí, P. & Turco, M. (2018). Use of bias correction techniques to improve seasonal forecasts for reservoirs — a case-study in northwestern mediterranean. *Science of The Total Environment*, **610-611**, 64–74. 14

Masson, V., Lemonsu, A., Hidalgo, J. & Voogt, J. (2020). Urban climates and climate change. *Annual Review of Environment and Resources*, **45**, 411–444. 15

McPhaden, M.J., Zebiak, S.E. & Glantz, M.H. (2006). Enso as an integrating concept in earth science. *Science*, **314**, 1740–1745. 5

Meehl, G.A., Goddard, L., Murphy, J., Stouffer, R.J., Boer, G., Danabasoglu, G., Dixon, K., Giorgetta, M.A., Greene, A.M., Hawkins, E., Hegerl, G., Karoly, D., Keenlyside, N., Kimoto, M., Kirtman, B., Navarra, A., Pulwarty, R., Smith, D., Stammer, D. & Stockdale, T. (2009). Decadal prediction: Can it be skillful? *Bulletin of the American Meteorological Society*. 7, 9

Merryfield, W.J., Baehr, J., Batté, L., Becker, E.J., Butler, A.H., Coelho, C.A., Danabasoglu, G., Dirmeyer, P.A., Doblas-Reyes, F.J., Domeisen, D.I., Ferranti, L., Ilynia, T., Kumar, A., Müller, W.A., Rixen, M., Robertson, A.W., Smith, D.M., Takaya, Y., Tuma, M., Vitart, F., White, C.J., Alvarez, M.S., Ardilouze, C., Attard, H., Baggett, C., Balmaseda, M.A., Beraki, A.F., Bhattacharjee, P.S., Bilbao, R., Andrade, F.M.D., DeFlorio, M.J., Díaz, L.B., Ehsan, M.A., Fragkoulidis, G., Grainger, S., Green, B.W., Hell, M.C., Infanti, J.M., Isensee, K., Kataoka, T., Kirtman, B.P., Klingaman, N.P., Lee, J.Y., Mayer, K., McKay, R., Mecking, J.V., Miller, D.E., Neddermann, N., Ng, C.H.J., Ossó, A., Pankatz, K., Peatman, S., Pegion, K., Perlwitz, J., Recalde-Coronel, G.C., Reintges, A., Renkl, C., Solaraju-Murali, B., Spring, A., Stan, C., Sun, Y.Q., Tozer, C.R., Vigaud, N., Woolnough, S. & Yeager, S. (2020). Current and emerging developments in subseasonal to decadal prediction. *Bulletin of the American Meteorological Society*, **101**, E869–E896. 4, 8, 10, 15

Mishra, N., Prodhomme, C. & Guemas, V. (2018). Multi-model skill assessment of seasonal temperature and precipitation forecasts over europe. *Climate Dynamics 2018 52:7*, **52**, 4207–4225. 14, 74

Murphy, A.H. (1992). Climatology, persistence, and their linear combination as standards of reference in skill scores. *Weather and Forecasting*, **7**, 692 – 698. 12, 74

Murphy, A.H. (1993). What is a good forecast? an essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, **8**, 281 – 293. 75

Müller, W.A., Jungclaus, J.H., Mauritsen, T., Baehr, J., Bittner, M., Budich, R., Bunzel, F., Esch, M., Ghosh, R., Haak, H., Ilyina, T., Kleine, T., Kornblueh, L., Li, H., Modali, K., Notz, D., Pohlmann, H., Roeckner, E., Stemmler, I., Tian, F. & Marotzke, J. (2018). A higher-resolution version of the max planck institute earth system model (mpi-esm1.2-hr). *Journal of Advances in Modeling Earth Systems*, **10**, 1383–1413. 10

O'Neill, B.C., Carter, T.R., Ebi, K., Harrison, P.A., Kemp-Benedict, E., Kok, K., Kriegler, E., Preston, B.L., Riahi, K., Sillmann, J. *et al.* (2020). Achievements and needs for the climate change scenario framework. *Nature climate change*, **10**, 1074–1084. 7

Pasternack, A., Bhend, J., Liniger, M.A., Rust, H.W., Müller, W.A. & Ulbrich, U. (2018). Parametric decadal climate forecast recalibration (deforest 1.0). *Geoscientific Model Development*, **11**, 351–368. 14

Paxian, A., Ziese, M., Kreienkamp, F., Pankatz, K., Brand, S., Pasternack, A., Pohlmann, H., Modali, K. & Früh, B. (2019). User-oriented global predictions of the gpcc drought index for the next decade. *Meteorologische Zeitschrift*, 3–21. 15, 76

Paxian, A., Reinhardt, K., Pankatz, K., Pasternack, A., Lorza-Villegas, M.P., Scheibel, M., Hoff, A., Mannig, B., Lorenz, P. & Früh, B. (2022). High-resolution decadal drought predictions for german water boards: A case study for the wupper catchment. *Frontiers in Climate*, **4**. 14

Pecl, G.T., Araújo, M.B., Bell, J.D., Blanchard, J., Bonebrake, T.C., Chen, I.C., Clark, T.D., Colwell, R.K., Danielsen, F., Evengård, B., Falconi, L., Ferrier, S., Frusher, S., Garcia, R.A., Griffis, R.B., Hobday, A.J., Janion-Scheepers, C., Jarzyna, M.A., Jennings, S., Lenoir, J., Linnetved, H.I., Martin, V.Y., McCormack, P.C., McDonald, J., Mitchell, N.J., Mustonen, T., Pandolfi, J.M., Pettorelli, N., Popova, E., Robinson, S.A., Scheffers, B.R., Shaw, J.D., Sorte, C.J.B., Strugnell, J.M., Sunday, J.M., Tuanmu, M.N., Vergés, A., Villanueva, C., Wernberg, T., Wapstra, E. & Williams, S.E. (2017). Biodiversity redistribution under climate change: Impacts on ecosystems and human well-being. *Science*, **355**, eaai9214. 1

Pohlmann, H., Kröger, J., Greatbatch, R.J. & Müller, W.A. (2017). Initialization shock in decadal hindcasts due to errors in wind stress over the tropical pacific. *Climate Dynamics*. 10

Pohlmann, H., Brune, S., Fröhlich, K., Jungclaus, J.H., Sgoff, C. & Baehr, J. (2023). Impact of ocean data assimilation on climate predictions with icon-esm. *Climate Dynamics*, **61**, 357–373. 10

Polkova, I., Brune, S., Kadow, C., Romanova, V., Gollan, G., Baehr, J., Glowienka-Hense, R., Greatbatch, R.J., Hense, A., Illing, S., Köhl, A., Kröger, J., Müller, W.A., Pankatz, K. & Stammer, D. (2019). Initialization and ensemble generation for decadal climate predictions: A comparison of different methods. *Journal of Advances in Modeling Earth Systems*, **11**, 149–172. 9

Portmann, F.T., Siebert, S. & Döll, P. (2010). Mirca2000—global monthly irrigated and rainfed crop areas around the year 2000: A new high-resolution data set for agricultural and hydrological modeling. *Global Biogeochemical Cycles*, **24**, GB1011. 78

Pérez-Zanón, N., Caron, L.P., Terzago, S., Schaeybroeck, B.V., Lledó, L., Manubens, N., Roulin, E., Alvarez-Castro, M.C., Batté, L., Delgado-Torres, C., Dom{\'i}nguez, M., von Hardenberg, J., Sánchez-Garc{\'i}a, E., Torralba, V. & Verfaillie, D. (2021). The cstools (v4.0) toolbox: from climate forecasts to climate forecast information. *Geoscientific Model Development Discussions*, **2021**, 1–32. 14

Riahi, K., van Vuuren, D.P., Kriegler, E., Edmonds, J., O'Neill, B.C., Fujimori, S., Bauer, N., Calvin, K., Dellink, R., Fricko, O., Lutz, W., Popp, A., Cuaresma, J.C., KC, S., Leimbach, M., Jiang, L., Kram, T., Rao, S., Emmerling, J., Ebi, K., Hasegawa, T., Havlik, P., Humpenöder, F., Da Silva, L.A., Smith, S., Stehfest, E., Bosetti, V., Eom, J., Gernaat, D., Masui, T., Rogelj, J., Strefler, J., Drouet, L., Krey, V., Luderer, G., Harmsen, M., Takahashi, K., Baumstark, L., Doelman, J.C., Kainuma, M., Klimont, Z., Marangoni, G., Lotze-Campen, H., Obersteiner, M., Tabeau, A. & Tavoni, M. (2017). The shared socioeconomic pathways and their energy, land use, and greenhouse gas emissions implications: An overview. *Global Environmental Change*, **42**, 153–168. 7

Rodó, X. & Comín, F.A. (2003). Global climate: Current research and uncertainties in the climate system. *Springer*. 1, 2

Saurral, R.I., García-Serrano, J., Doblas-Reyes, F.J., Díaz, L.B. & Vera, C.S. (2020). Decadal predictability and prediction skill of sea surface temperatures in the south pacific region. *Climate Dynamics*, **54**, 3945–3958. 5

Scaife, A.A. & Smith, D. (2018). A signal-to-noise paradox in climate science. *npj Climate and Atmospheric Science*, **1**, 28. 11

Schaeybroeck, B.V. & Vannitsem, S. (2011). Post-processing through linear regression. *Nonlinear Processes in Geophysics*, **18**, 147–160. 14

Schaeybroeck, B.V. & Vannitsem, S. (2015). Ensemble post-processing using member-by-member approaches: theoretical aspects. *Quarterly Journal of the Royal Meteorological Society*, **141**, 807–818. 14

Schuster, M., Grieger, J., Richling, A., Schartner, T., Illing, S., Kadow, C., Müller, W.A., Pohlmann, H., Pfahl, S. & Ulbrich, U. (2019). Improvement in the decadal prediction skill of the north atlantic extratropical winter circulation through increased model resolution. *Earth System Dynamics*, **10**, 901–917. 10

Sillmann, J., Thorarinsdottir, T., Keenlyside, N., Schaller, N., Alexander, L.V., Hegerl, G., Seneviratne, S.I., Vautard, R., Zhang, X. &

Zwiers, F.W. (2017). Understanding, modeling and predicting weather and climate extremes: Challenges and opportunities. *Weather and Climate Extremes*, **18**, 65–74. 1

Smith, D.M., Cusack, S., Colman, A.W., Folland, C.K., Harris, G.R. & Murphy, J.M. (2007). Improved surface temperature prediction for the coming decade from a global climate model. *Science*, **317**, 796–799. 7

Smith, D.M., Eade, R. & Pohlmann, H. (2013). A comparison of full-field and anomaly initialization for seasonal to decadal climate prediction. *Climate Dynamics 2013 41:11*, **41**, 3325–3338. 9, 14

Smith, D.M., Booth, B.B., Dunstone, N.J., Eade, R., Hermanson, L., Jones, G.S., Scaife, A.A., Sheen, K.L. & Thompson, V. (2016). Role of volcanic and anthropogenic aerosols in the recent global surface warming slowdown. *Nature Climate Change*, **6**, 936–940. 3

Smith, D.M., Eade, R., Scaife, A.A., Caron, L.P., Danabasoglu, G., DelSole, T.M., Delworth, T., Doblas-Reyes, F.J., Dunstone, N.J., Hermanson, L., Kharin, V., Kimoto, M., Merryfield, W.J., Mochizuki, T., Müller, W.A., Pohlmann, H., Yeager, S. & Yang, X. (2019). Robust skill of decadal climate predictions. *npj Climate and Atmospheric Science 2019 2:1*, **2**, 1–10. 9

Smith, D.M., Scaife, A.A., Eade, R., Athanasiadis, P., Bellucci, A., Bethke, I., Bilbao, R., Borchert, L.F., Caron, L.P., Counillon, F., Danabasoglu, G., Delworth, T., Doblas-Reyes, F.J., Dunstone, N.J., Estella-Perez, V., Flavoni, S., Hermanson, L., Keenlyside, N., Kharin, V., Kimoto, M., Merryfield, W.J., Mignot, J., Mochizuki, T., Modali, K., Monerie, P.A., Müller, W.A., Nicolí, D., Ortega, P., Pankatz, K., Pohlmann, H., Robson, J., Ruggieri, P., Sospedra-Alfonso, R., Swingedouw, D., Wang, Y., Wild, S., Yeager, S., Yang, X. & Zhang, L. (2020). North atlantic climate far more predictable than models imply. *Nature 2020 583:7818*, **583**, 796–800. 11, 12, 14, 86

Solaraju-Murali, B. (2023). On the use of decadal predictions for agricultural climate services: bridging the gap between service providers and users. *Ph.D. thesis, Universitat Politècnica de Catalunya, Barcelona*. 15, 18, 76

Solaraju-Murali, B., Gonzalez-Reviriego, N., Caron, L.P., Ceglar, A., Toreti, A., Zampieri, M., Bretonnière, P.A., Cabré, M.S. & Doblas-Reyes, F.J. (2021). Multi-annual prediction of drought and heat stress to support decision making in the wheat sector. *npj Climate and Atmospheric Science 2021 4:1*, **4**, 1–9. 15, 18

Solaraju-Murali, B., Bojovic, D., Gonzalez-Reviriego, N., Nicodemou, A., Terrado, M., Caron, L.P. & Doblas-Reyes, F.J. (2022). How decadal predictions entered the climate services arena: an example from the agriculture sector. *Climate Services*, **27**, 100303. 11, 15

Sospedra-Alfonso, R. & Boer, G.J. (2020). Assessing the impact of initialization on decadal prediction skill. *Geophysical Research Letters*, **47**, e2019GL086361. 8

Talman, C.F. (1927). The early days of weather forecasting. *Bulletin of the American Meteorological Society*, **8**, 147–148. 6

Thompson, D.W.J. & Wallace, J.M. (1998). The arctic oscillation signature in the wintertime geopotential height and temperature fields. *Geophysical Research Letters*, **25**, 1297–1300. 3

Tippett, M.K. & Barnston, A.G. (2008). Skill of multimodel enso probability forecasts. *Monthly Weather Review*, **136**, 3933–3946. 14

Tommasi, D., Stock, C.A., Hobday, A.J., Methot, R., Kaplan, I.C., Eveson, J.P., Holsman, K., Miller, T.J., Gaichas, S., Gehlen, M., Pershing, A., Vecchi, G.A., Msadek, R., Delworth, T., Eakin, C.M., Haltuch, M.A., Séférian, R., Spillman, C.M., Hartog, J.R., Siedlecki, S., Samhouri, J.F., Muhling, B., Asch, R.G., Pinsky, M.L., Saba, V.S., Kapnick, S.B., Gaitan, C.F., Rykaczewski, R.R., Alexander, M.A., Xue, Y., Pegion, K.V., Lynch, P., Payne, M.R., Kristiansen, T., Lehodey, P. & Werner, F.E. (2017). Managing living marine resources in a dynamic environment: The role of seasonal to decadal climate forecasts. *Progress in Oceanography*, **152**, 15–49. 15

Torralba, V. (2019). Seasonal climate prediction for the wind energy sector: methods and tools for the development of a climate service. *Ph.D. thesis, Universidad Complutense de Madrid*. 18

Torralba, V., Doblas-Reyes, F.J., MacLeod, D., Christel, I. & Davis, M. (2017). Seasonal climate prediction: A new source of information for the management of wind energy resources. *Journal of Applied Meteorology and Climatology*, **56**, 1231–1247. 13, 14

Trascasa-Castro, P., Ruprich-Robert, Y., Castruccio, F. & Maycock, A.C. (2021). Warm phase of amv damps enso through weakened thermocline feedback. *Geophysical Research Letters*, **48**, e2021GL096149, e2021GL096149 2021GL096149. 3

Trenberth, K.E. & Shea, D.J. (2006). Atlantic hurricanes and natural variability in 2005. *Geophysical Research Letters*, **33**, 12704. 5

Verfaillie, D., Doblas-Reyes, F.J., Donat, M.G., Pérez-Zanón, N., Solaraju-Murali, B., Torralba, V. & Wild, S. (2021). How reliable are decadal climate predictions of near-surface air temperature? *Journal of Climate*, **34**, 697 – 713. 12

Vicente-Serrano, S.M., Beguería, S. & López-Moreno, J.I. (2010). A multiscalar drought index sensitive to global warming: The standardized precipitation evapotranspiration index. *Journal of Climate*, **23**, 1696 – 1718. 16

Visbeck, M.H., Hurrell, J.W., Polvani, L. & Cullen, H.M. (2001). The north atlantic oscillation: Past, present, and future. *Proceedings of the National Academy of Sciences*, **98**, 12876–12877. 3

Vitart, F. & Robertson, A.W. (2018). The sub-seasonal to seasonal prediction project (s2s) and the prediction of extreme events. *npj Climate and Atmospheric Science*, **1**, 3. 7

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Déqué, M., Ferranti, L., Fucile, E., Fuentes, M., Hendon, H., Hodgson, J., Kang, H.S., Kumar, A., Lin, H., Liu, G., Liu, X., Malguzzi, P., Mallas, I., Manoussakis, M., Mastrangelo, D., MacLachlan, C., McLean, P., Minami, A., Mladek, R., Nakazawa, T., Najm, S., Nie, Y., Rixen, M., Robertson, A.W., Ruti, P., Sun, C., Takaya, Y., Tolstykh, M., Venuti, F., Waliser, D., Woolnough, S., Wu, T., Won, D.J., Xiao, H., Zaripov, R. & Zhang, L. (2017). The subseasonal to seasonal (s2s) prediction project database. *Bulletin of the American Meteorological Society*, **98**, 163 – 173. 7

Volpi, D., Meccia, V.L., Guemas, V., Ortega, P., Bilbao, R., Doblas-Reyes, F.J., Amaral, A., Echevarria, P., Mahmood, R. & Corti, S. (2021). A novel initialization technique for decadal climate predictions. *Frontiers in Climate*, **3**. 9, 10

Ward, D.S., Mahowald, N.M. & Kloster, S. (2014). Potential climate forcing of land use and land cover change. *Atmospheric Chemistry and Physics*, **14**, 12701–12724. 6

Weigel, A.P., Liniger, M.A. & Appenzeller, C. (2009). Seasonal ensemble forecasts: Are recalibrated single models better than multimodels? *Monthly Weather Review*, **137**, 1460–1479. 14

Weisheimer, A. & Palmer, T.N. (2014). On the reliability of seasonal climate forecasts. *Journal of the Royal Society Interface*, **11**, 20131162. 13

Wilks, D.S. (2011). Forecast verification. *International Geophysics*, **100**, 301–394. 74

Woolnough, S.J., Vitart, F. & Balmaseda, M.A. (2007). The role of the ocean in the madden–julian oscillation: Implications for mjo prediction. *Quarterly Journal of the Royal Meteorological Society*, **133**, 117–128. 3

Zanchettin, D. (2017). Aerosol and solar irradiance effects on decadal climate variability and predictability. *Current Climate Change Reports*, **3**, 150–162. 6

Zhang, X., Alexander, L., Hegerl, G.C., Jones, P., Tank, A.K., Peterson, T.C., Trewin, B. & Zwiers, F.W. (2011). Indices for monitoring changes in extremes based on daily temperature and precipitation data. *Wiley Interdisciplinary Reviews: Climate Change*, **2**, 851–870. 16, 76

Zhao, T., Bennett, J.C., Wang, Q.J., Schepen, A., Wood, A.W., Robertson, D.E. & Ramos, M.H. (2017). How suitable is quantile mapping for postprocessing gcm precipitation forecasts? *Journal of Climate*, **30**, 3185–3196. 14

Zumwald, M., Knüsel, B., Baumberger, C., Hirsch Hadorn, G., Bresch, D.N. & Knutti, R. (2020). Understanding and assessing uncertainty of observational climate datasets for model evaluation using ensembles. *WIREs Climate Change*, **11**, e654. 9

Decadal climate prediction and predictability for climate services

Author
Carlos Delgado-Torres

A thesis submitted for the degree of Doctor of Philosophy

Climate variations at annual to decadal time scales impact the economy, ecosystems and society in several sectors, such as renewable energy, agriculture, food security, water management, fisheries, health, insurance and urban planning. Knowing these variations ahead of time allows for implementing measures to adapt, mitigate and build resilience to the consequences of a changing climate. The work developed within this Ph.D. thesis has focused on evaluating the forecast quality for predictions of several variables, indices and indicators relevant for decision-making in several sectors, with a particular focus on agriculture. The evaluation has been performed globally, for the individual models and multi-model ensemble, and different forecast periods in order to identify windows of opportunity for which the climate predictions show enough quality to be used for decision-making. Besides, different post-processing techniques have been applied to the predictions to improve their quality and usability. The thesis also presents some applications of the research within different projects, as well as prototypes of climate services developed in collaboration with users.