npj Climate and Atmospheric Science https://doi.org/10.1038/s41612-026-01320-z

Article in Press

Data-driven seasonal climate predictions via variational inference and transformers

Received: 28 February 2025

Accepted: 1 January 2026

Cite this article as: Palma, L., Peraza, A., Civantos-Prieto, D. *et al.* Data-driven seasonal climate predictions via variational inference and transformers. *npj Clim Atmos Sci* (2026). https://doi.org/10.1038/s41612-026-01320-z

Lluís Palma, Alejandro Peraza, David Civantos-Prieto, Amanda Duarte, Stefano Materia, Ángel G. Muñoz, Jesús Peña-Izquierdo, Laia Romero, Albert Soret & Markus G. Donat

Data-driven Seasonal Climate Predictions via Variational Inference and Transformers

Lluís Palma[1] [2]✉, Alejandro Peraza[1], David Civantos-Prieto[3], Amanda Duarte[1], Stefano Materia[1], Ángel G. Muñoz[1], Jesús Peña-Izquierdo[3], Laia Romero[3], Albert Soret[1], and Markus G. Donat[1] [4]

[1] Barcelona Supercomputing Center, Earth Sciences Department, Barcelona, 08034, Spain

[2] Facultat de Física, Universitat de Barcelona, Barcelona, 08028, Spain

[3] Lobelia Earth, Barcelona, 08005, Spain

[4] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, 08005, Spain

✉ Corresponding author: Lluís Palma (lluis.palma@bsc.es)

Abstract

Most operational climate services providers base their seasonal predictions on initialised general circulation models (GCMs) or empirical statistical techniques. GCMs are widely used but require substantial computational resources, limiting their capacity. In contrast, statistical methods often lack robustness due to the short historical records available. Recent works propose machine learning methods trained on climate model output, leveraging larger sample sizes. Yet, many of these studies focus on prediction tasks that may be restricted in spatial or temporal extent, thereby creating a gap with existing operational predictions. Others fail to disentangle the sources of skill in the context of climate change, where strong trends provide spurious estimates. This study combines variational inference with transformers to predict global and regional seasonal anomalies of temperature and rainfall. The model is trained on output from CMIP6 and tested using ERA5 reanalysis data. Temperature predictions demonstrate skill beyond the climatology and climate-change trend and even outperform the numerical state-of-the-art system SEAS5 in some ocean and land areas. Precipitation forecasts show more limited skill, with fewer regions outperforming climatology and fewer surpassing SEAS5. Furthermore, the consistency found in both teleconnections and skill spatial patterns against SEAS5 suggests that both systems build on similar sources of predictability.

Introduction

In contrast to weather forecasts, which predict daily atmospheric conditions for up to two weeks, seasonal climate predictions provide estimates of seasonal statistics months in advance. To address the inherent unpredictability of the atmosphere[1] and the resulting stochasticity of the Earth system, seasonal climate predictions leverage ocean and land surface forcings in conjunction with ensemble predictions that provide probabilistic information[2]. Thus, seasonal outlooks commonly deliver probabilities of wetter/drier or warmer/colder than average conditions. This information has proven valuable for many climate-sensitive sectors, including agriculture[3,4], renewable energy production[5–7] or public health[8]. Consequently, over the past few decades, seasonal climate prediction has transformed from a research effort into an operational service[9,10]. Nonetheless, the practical value of seasonal predictions relies on their skill and resolution, and for many applications, those might not reach user requirements[11].

Current operational climate services providers base their seasonal predictions on dynamical models, statistical methods, or a hybrid combination. Dynamical models are based on coupled (atmosphere, land, ocean, and sea-ice) General Circulation Models (GCMs), which embody the most complete representation of climate system dynamics known. These models are typically initialised to our best estimate of the observed climate state[12], integrating the diverse mix of observations available into a consistent set of fields through a process known as data assimilation. Those fields provide an initial state of the Earth system and serve as a starting point for a simulation covering the desired prediction period, typically up to eight months for seasonal predictions. Simulating the Earth system at a global scale over long periods and multiple ensemble members requires vast computational resources, limiting the model's spatial resolution[13]. Such limitation cascades into many physical processes not being explicitly resolved and subject to parametrization, resulting in biased dynamics. These biases add to an imperfect definition of the initial state, partly due to an erratic spatial and temporal distribution of current observations[14,15]. This combination of factors results in seasonal predictions presenting strong drifts and biases, deteriorating prediction quality[11,16].

On the other hand, statistical or empirical prediction methods benefit from efficiently leveraging the relationships learned from past observational records. They explicitly capture the interactions between predictors and predictands, offering a more direct approach than dynamical models, which derive such relationships through iterative simulation at finer temporal scales. Statistical methods range from simple persistence models to sophisticated statistical techniques[17], including machine or deep learning algorithms[18]. These approaches can yield skilful forecasts comparable to dynamical models. Still, statistical models are not exempt from errors and require careful application due to short observational records and climate non-stationarity, which often compromises the independent and identically distributed (i.i.d.) assumption, paramount for many of these methods.

The limited temporal extent of current observational datasets poses a greater challenge when involving large machine learning (ML) algorithms. Training these algorithms with a limited dataset results in almost certain overfitting due to the imbalance between trainable parameters and available training samples. Unlike weather forecasting, where high-frequency temporal variability yields multiple independent samples over short periods, seasonal processes operate on monthly to annual scales. At these timescales, interannual processes dominate, resulting in as few as one independent sample per year. Consequently, seasonal forecasting applications have far fewer data points, up to two orders of magnitude less than weather applications trained over the same period[19–21]. This limitation partly explains the explosion in studies applying large deep learning models to weather forecasting[22–26], contrasting with the few works tackling seasonal climate predictions.

To address such limitations, most ML applications for seasonal predictions rely on climate model output to train their data-driven models[20,27,28]. The underlying assumption is that climate models can, to a certain extent, simulate the climate system and its interannual variability, making the thousands of simulated years a valid training set. Beyond deep learning, this logic has also motivated works in the climate community where analogues from climate model output assemble seasonal to decadal predictions from initial conditions[29–32]. Some even employ deep learning to select the best analogues[33]. Training with climate model output effectively increases the number of samples from tens to thousands when using multiple simulations. This approach

also delivers data from various multi-decadal periods, as the simulations cover hundreds of years, and allows learning from unobserved regime shifts or trends, such as the one forced by global warming. Data that captures regime shifts and unseen scenarios is crucial for most ML algorithms, which typically struggle with extrapolation and are highly sensitive to regime shifts due to their reliance on the i.i.d. assumption[34,35]. However, as previously mentioned, predictions based on numerical climate models (GCMs) present known drifts and biases and misrepresent or fail to resolve critical physical processes, resulting in poor simulation of some key teleconnections that contribute to predictability. Consequently, ML models trained on climate model output are potentially limited by the climate model's ability to simulate the processes relevant to the seasonal prediction task of interest.

Current applications of deep learning algorithms trained on climate model output include the prediction of El Nino-Southern Oscillation (ENSO) and the Indian Ocean Dipole (IOD)[27,36], the prediction of Arctic sea-ice[19], the occurrence of drought events in the US[20], or the prediction of European summer heatwaves[37] and droughts[28]. Most studies rely on classical machine learning methods or simpler deterministic neural networks and focus on specific prediction tasks that tend to collapse spatial information for simplicity. Providing global probabilistic predictions with data-driven models is essential for a robust comparison against current state-of-the-art dynamical prediction systems[17]. As an example, this study[38] uses a probabilistic method to predict (Oct-Mar) precipitation and temperature anomalies based on the previous July's upper ocean thermal status. However, how skill varies across seasons remains to be explored, especially at lead times closer to the initial state, which is of particular interest to users. In addition to that, in the context of global warming where strong trends provide added predictability on top of interannual fluctuations[39,40], it is essential to disentangle both sources of predictability to properly assess the added value of the tested prediction systems[11,41,42].

This study evaluates the effectiveness of an intrinsically probabilistic deep learning method for predicting global three-month seasonal anomalies (lead months 1-3) of temperature and precipitation fields throughout the year. The model is trained using the output of CMIP6 and validated against ERA5 reanalysis data. It explicitly decomposes the contribution of climate change-induced trends during both training and validation. Thus, the approach combines

variational inference with vision transformers to explicitly predict interannual seasonal anomalies. To the best of our knowledge, this is the first application of vision transformers for seasonal prediction. Additionally, we apply this methodology in a regional context, specifically focusing on Europe, where we compare the effects and robustness of targeting different spatial domains and resolutions. Finally, to gain understanding of the sources of predictability, we also assess the skill and teleconnection patterns that emerge from two primary modes of variability at the seasonal timescale: the El Niño-Southern Oscillation (ENSO) and the North Atlantic Oscillation (NAO).

Results

Forecasts assessment

The presented approach (illustrated in Figure 1) uses a conditional Variational Autoencoder (cVAE [43,44]) architecture to predict seasonal climate anomalies. The model takes monthly means of five essential climate variables from the preceding six months as input: 2-meter air temperature (tas), precipitation (pr), sea surface temperature (tos), and geopotential height at 500hPa and 300hPa levels (zg500, zg300), and predicts their 3-month seasonal averages, on lead months 1 to 3 (i.e., for a prediction that leverages information up to November 1st, a seasonal mean for DJF is predicted). Two vision transformers[45] encode the input and target states into a latent space, capturing the underlying climate patterns. Transformers are based on a general-purpose inductive bias that separates the interaction range from the network's depth. This separation enables the modelling of both distant and local connections without requiring a complex hierarchy of convolutional neural network (CNN)[46] operations. As a result, Vision Transformers (ViTs) can more explicitly capture both local and long-range climate interactions. To decode and generate the probabilistic predictions, we employ a CNN that processes multiple samples from the latent space to create an ensemble forecast. The model explicitly predicts the interannual variability component, while a separate locally estimated scatterplot smoothing (LOESS) regression handles the long-term trend. The output of these components is later combined to produce the final forecast. This setup represents an example of using simple statistical methods to capture predictable signals while leveraging neural networks to model complex deviations. The model

was trained using Coupled Model Intercomparison Project Phase 6 (CMIP6)[47] data and tested against the European Centre for Medium-Range Weather Forecasts Reanalysis v5 (ERA5)[48] dataset. Further details about the methodology can be found in the Methods section.

Examples of predicted global temperature (Figures S1 and S2) and precipitation (Figures S3 and S4) anomalies are provided in the supplementary information. Figures S1 and S3 show the anomalies of each individual ensemble member for the 2020 SON prediction, while Figures S2 and S4 show anomalies' ensemble medians for the 2002-2021 period (also SON). Variability among members is minimal over oceans and equatorial areas, where seasonal predictability is strongest, while inter-ensemble variability over land and in the extratropics is much more pronounced. The ensemble medians across different years reveal the increasing effects of global warming on temperature predictions, and the fingerprint of El Niño (2002, 2009, 2015, 2020) and La Niña years (2007, 2010) is evident in global sea surface temperatures and rainfall patterns.

Seasonal climate predictions are influenced by multiple processes operating at different time scales with varying degrees of predictability. Predictability at the seasonal time scale is affected not only by interannual fluctuations but also by longer-term modulations such as trends driven by greenhouse gas emissions or lower-frequency decadal oscillations. These longer-term processes may have different levels of predictability at the interannual scale. However, seasonal predictions aim to provide information on seasonal anomalies at the interannual level, going beyond trends or decadal oscillations. Thus, disentangling all these different signals in our validation procedure is crucial to understanding the value of the seasonal predictions, if any.

Figure 2 shows the Anomaly Correlation Coefficient (ACC), computed between our predictions and ERA5 reanalysis for both de-trended temperature (trended results are shown in Figure S11) and precipitation seasonal anomalies. The predictions were initialized one month before the start of the season, i.e. the DJF prediction used climate information up to November 1st. Overall, the ACC validation against ERA5 reveals a generally higher skill for temperature predictions compared to precipitation, with most of the skill concentrated in the tropics and limited skill in the extratropical regions. Both outcomes are in line with numerical forecasts[49,50], suggesting that the

signal to noise ratio is not altered in the cVAE prediction, and that the ML model taps into similar underlying teleconnections.

Temperature forecasts exhibit stronger correlations over oceanic regions, with particularly robust signals in the equatorial Pacific across seasons, peaking in the SON and DJF seasons. Beyond the ENSO signature in the equatorial Pacific is the Pacific Decal Oscillation (PDO) signature that extends into the extra-tropical north-Pacific Sea. Significant correlations in temperature are observed over land areas in Central America, Brazil, Australia, central and north (DJF) Africa, South Africa (DJF), and southeast central Asia. Although more limited, Europe also shows positive correlations (MAM, SON & DJF). Precipitation forecasts, are generally less skilful, positive ACC patterns are spatially less extent compared temperature predictions, but still show high correlations (above 0.7) in Indonesia (SON, DJF, MAM) and the Caribbean (JJA), with moderate correlations (0.5-0.7) in northern South America (SON, DJF), Australia (SON), the Horn of Africa (SON), and the US (SON, DJF & JJA). Weaker (0.3-0.5) precipitation correlations are observed in similar regions for other seasons, including parts of Europe, India, the central Atlantic coast of Africa, and southern South America. Results for an extended period 1985-2021 are shown in the Supplementary information (Figures S11, S12 & S14). These results are not directly assessed in the main manuscript as the time period employed overlaps with the reference period defined to compute the standardization and de-trending[51]. Overall, our method captures predictable signals that are physically realistic and consistent with findings reported in the literature[11,17], demonstrating the potential of the methodology for seasonal forecasting applications.

The generative nature of variational methods enables the production of multiple deterministic predictions, forming an ensemble of plausible outcomes. This ensemble facilitates the derivation of a more robust deterministic signal. Inferencing a complete hindcast (1950-2021), 100 members, for all variables and a single season takes seconds on a single H100 GPU. We make an initial assessment of the ensemble and the impact of its size on the model's performance, shown in Figure 3. We observe that increasing the ensemble size from 1 to 100 members enhances forecast skill, with the ACC improving from 0.07 to 0.51. The model reaches near-optimal performance with an ACC of 0.49 using just 20 ensemble members, indicating a balance between computational efficiency and forecast accuracy. Larger ensemble sizes provide a more explicit

representation of forecast uncertainty, while individual members reflect realistic temporal variability, closely following observed climate variability. The model effectively captures key aspects of ensemble forecasting systems while offering substantial computational benefits. The ability of generating diverse ensembles while obtaining peak correlations with moderate ensemble sizes are desirable properties for climate prediction.

Validation against benchmarks

Supplementary Figures S9 to S14 show differences in the anomaly correlation coefficient (ACC) between SEAS5 and our approach. However, we acknowledge that ACC differences can be problematic due to the bounded nature of correlations (-1 to 1) and the resulting skewed sampling distribution[52], which enforces the use of Fisher's z-transform for proper statistical analysis. Additionally, correlation measures only capture linear associations between forecasts and observations, neglecting important factors such as biases and non-linear relationships. For these reasons, we prefer the use of skill scores in our benchmark comparisons. Nonetheless, readers interested in an ACC analysis can refer to supplementary Figures S9 to S14, which produce qualitatively similar conclusions to those obtained using skill scores, considering the differences between the metrics.

We compare our model with the climatological forecast (CLIM) and the ECMWF's seasonal prediction system (SEAS5). The period covering 1981-2000 has been used as the reference period for the climatology and anomalies. As stated in the previous section, lead months 1 to 3 are considered, which is equivalent to taking the SEAS5 prediction initialised on November 1st, targeting DJF. Figure 4 shows the forecast skill scores for near-surface air temperature (tas) predictions from 2001-2021. We present the results for four seasons (DJF, MAM, JJA, SON) across columns and using two skill metrics: the root-mean-square error skill score (RMSS) and the continuous ranked probability skill score (CRPSS). While the RMSS measures the deterministic performance of the ensemble median, the CRPSS offers a better view of the probabilistic performance of the ensemble distribution. Skill scores range from -2 (dark pink), where negative values indicate no skill, to 0.5 (dark blue), indicating high skill above the reference. Figure 5 shows the results of precipitation predictions using the same two metrics. Black dots indicate whether

the positive skills are statistically significant at the 95% confidence interval (more details in Methods). Again, results for an extended period 1985-2021 and without de-trending are shown in the Supplementary information (Figures S15-18).

Overall, the skill metrics relative to the climatological forecast exhibit similar patterns to those in Figure 2, showing consistent performance across seasons. Temperature fields demonstrate higher and more spatially extensive skill compared to precipitation predictions. Similarly, for temperature, we observe a predominately more robust signal over the oceans compared to land regions and a clear ENSO and PDO signature. Precipitation forecasts also show a similar signal compared to the one in Figure 2. We find strong performance (Skill score above 0.2) compared to climatology in regions such as northern South America (DJF & MAM), Australia (MAM & SON), eastern (DJF & MAM) and southern Africa (DJF), as well as parts of the U.S. (DJF, MAM & SON) and the Arctic (DJF & SON). However, performance varies by season and region, with limited skill in Europe.

We find fewer regions showing improvements when comparing our model's performance against SEAS5 (third and fourth rows of Figures 4 and 5). This is expected, as SEAS5 is the current state-of-the-art dynamical forecasting system, making it a more challenging benchmark to surpass compared to climatology. Generally, we find that SEAS5 outperforms our approach in many oceanic areas. Nevertheless, our model performs better in some land regions, which are particularly relevant for user applications. Regarding temperature forecasts, our approach shows season-dependent improvements over SEAS5 in parts of the United States (MAM & JJA), the southern portion of South America (DJF, MAM & JJA), north Africa (MAM & DJF), Europe (JJA, SON & DJF), and Eurasia (particularly in the SON season). In contrast, our precipitation forecasts are clearly outperformed by SEAS5 in most regions. Yet, we find season-dependent enhancements in some extra-tropical regions, i.e., Europe (JJA), Eurasia (SON), parts of North America (SON), and the most southern part of South America (MAM & JJA).

It should be noted that the added value from our predictions is only realised when there is an improvement over both SEAS5 and the climatological forecast. If this improvement is not achieved jointly, our data-driven approach may converge to a simple climatological forecast. To

complement our results, we included in the Supplementary information Figures S19-24 a set of scorecards comparing CRPS values for the climatological forecast (clim), our approach (cVAE) and SEAS5. We computed the CRPS over the IPCC AR6 reference regions[53]. We split the results over land and ocean regions.

Overall, the climatology is rarely the best performer, validating the value of our system where it has higher skill compared to SEAS5. Similarly, we also find that skill exceeds the climatology mainly for temperature, and much less for precipitation, i.e., the lower signal to noise ratio in rainfall anomalies implies a less robust distinction (in terms of skill) between prediction systems. SEAS5 clearly outperforms our approach over oceanic regions, while on land areas the cVAE seems more competitive. More in detail, temperature forecasts (Figure S20) show seasonal-dependent skill improvements in parts of the United States (WNA, CNA, ENA, NCA), northwest Africa (SHA), the Arctic (GIC), and scattered locations in the rest of Africa (WAF, CAF, NEAF & SEAF) and South America (SES, SSA, SWS). The cVAE precipitation forecasts (Figure S22) are only improved in few extratropical regions, including small parts of Europe (WCE, MED, NEU), Eurasia (WSB, ESB), the US (NCA), and South America. However, such improvements are often marginal (Figure 5).

European regional use-case

Once we tested the approach globally and verified that the model reproduces well-known patterns at the seasonal time scale, we verified the model with a Europe-centric context. As shown in Figure 1 (panel D), three configurations are tested, where models with the same inputs X have three different targets Y: Global predictions at a spatial resolution of 5°, regional also at 5° and regional at 1°. The main objective is to assess whether targeting a more constrained region can yield improvements, as the model does not need to optimize its parameters for predicting the whole globe. Similarly, due to a smaller target domain, we can increase the spatial resolution of the target variable while maintaining the comparable computational cost needed for the original model (Global 5°). Thus, we assess whether we find benefits from increasing the spatial resolution of the predictions. Additionally, we test whether the patterns are consistent across the different configurations.

The supplementary information (Figures S5-S8) includes examples of predicted European temperature and precipitation patterns for the regional 1° configuration. Temperature forecasts reveal strong climate change forcing signals in both the individual ensemble members and the ensemble median, indicating that external forcing dominates over interannual signals in this region. Precipitation predictions exhibit north-south and east-west dipole patterns resembling those associated with different atmospheric regimes[54], such as the Arctic Oscillation (AO), the North Atlantic Oscillation (NAO), the Eastern Atlantic pattern (EA) or the Scandinavian blocking (SB). However, the significant variability among ensemble members highlights the model's difficulty in identifying a preferred pattern under these low-predictability conditions.

The CRPSS for temperature predictions over Europe (Figure 6) shows varying degrees of skill compared to both climatology and SEAS5. As a first assessment, we can see across the different models and seasons higher skill values against SEAS5 than the climatology, indicating the poor performance of SEAS5 in Europe. We also observe consistent patterns across model configurations, with the Regional 1° offering the most detailed spatial representation of skill, highlighting statistically significant improvements in specific regions while still aligning with patterns observed at coarser resolutions. Across the different configurations we observe how the model manages to improve the predictions over SEAS5 and the climatology for large parts of the Central and Western Mediterranean (across seasons), and some parts of Northern and Central Europe (MMA and JJA). As mentioned in the previous section, improvements over SEAS5 and the climatology are needed to add value to user-centred applications.

For precipitation forecasts, we find, in general terms (i.e., across seasons and regions), lower precipitation skills against the climatological forecasts than those obtained against SEAS5 (Figure 7), indicating the low performance in Europe of current state-of-the-art prediction systems. Yet, our approach surpasses SEAS5's skill slightly for different regions and seasons. We observe slight improvements both against the climatology and SEAS5 over parts of central to eastern Europe (DJF and JJA), parts of the British Isles and northern Europe. Again, we observe how the general patterns are maintained against the different configurations, with the 1° configuration highlighting a more detailed but coherent spatial representation of skill and significance. Again,

results for an extended period 1985-2021 are shown in the Supplementary information (Figures S25-26).

The robustness of the results across various spatial configurations (Global 5°, Regional 5°, and Regional 1°) represents an interesting finding. The skill patterns remain consistent regardless of the domain's resolution or setup. In addition, we observe that in most cases, the skill is either retained or improved by going regional or increasing the spatial resolution. Therefore, the choice of the most suitable configuration ultimately depends on user needs and computational constraints.

Teleconnections assessment

As a preliminary assessment of the sources of predictability in our model, we investigate the teleconnection patterns that emerge from two primary modes of variability at the seasonal timescale: the El Niño-Southern Oscillation (ENSO) and the North-Atlantic Oscillation (NAO). We compare such patterns against the ones found in SEAS5 and ERA5. For ENSO, we compute the temporal Pearson correlation (2001-2021) between the DJF SST Niño 3.4 index and DJF temperature, precipitation and 500hPa geopotential fields. For SEAS5 and our method (cVAE), the correlation is also computed along the ensemble dimension, i.e., the ensemble and time dimensions are concatenated into a grid-point-by-grid-point one-dimensional time series. A bootstrap procedure is applied with a fixed sample size to account for the varying dataset sizes, as forecasting systems often have a diverse number of ensemble members. The bootstrap approach also provides a range of uncertainty from which significance can be tested. Supplementary information Figure S27 shows a time series with the predicted ENSO index for the various systems against ERA5. We notice a slight reduction (0.96 cVAE vs 0.99 SEAS5) in the correlation against ERA5 and an overall wider uncertainty range for the generative model.

Examining the correlation maps in Figure 8, we find, in general terms, similar spatial patterns among the three datasets, with SEAS5 exhibiting weaker correlations (especially for precipitation) compared to ERA5 or the cVAE, which displays an over-amplified response. Linking these results with the CRPSS plots shown in Figures 4 & 5, we observe that for temperature fields, a

north/south dipole of positive/negative correlations over South America found in ERA5 is more closely represented in the cVAE compared against SEAS5, consistent with the CRPSS increase referenced against SEAS5. Similarly, higher correlations in the Horn of Africa and South Africa are found, consistent with ERA5 and the positive CRPSS against SEAS5 (whose correlation pattern is not aligned with ERA5). Regionally negative skill score values can also be explained by examining the weak to missing Niño3.4 temperature responses in the Indian Ocean, the equatorial Atlantic, and the PNA region (also noticeable in 2T and ZG500). Regarding precipitation fields, we also find several correlation patterns from the cVAE that are closer to those found in ERA5, leading to an increase in CRPSS. These include parts of central and south-east Africa, Southeastern South America and the Philippines. Yet, incoherent correlation patterns, between the cVAE and ERA5, such as those found in parts of East Indonesia and Central and North-East South America, are associated with significant negative CRPSS values.

A similar analysis is performed for the NAO. To define the NAO index, the 1st EOF of ERA5 500hPa fields over the North Atlantic sector (30°- 88.5°N, 80°W - 40°E) is used (Supplementary Figure S28). The 500hPa fields of the different systems are projected into such EOF, obtaining a NAO index for each system. Supplementary Figure S29 shows a time series with the predicted NAO index for the various systems against ERA5. We observe a slight increase (0.31 cVAE vs 0.22 SEAS5) in correlation against ERA5 and an overall similar uncertainty range. Regarding the correlation maps in Figure 9, we find, in general terms, a stronger signal captured by the cVAE compared against SEAS5. For temperature, a stronger dipole between the east coast of the US and the Labrador Sea, and similarly between the north of Europe and North Africa. More in detail, we find an extension of the correlation pattern into the Mediterranean and north/west of the black sea, consistent with skill improvements compared to SEAS5 and the climatology. Concerning rainfall, a tripole structure (observed in ERA5) between the Labrador Sea, Northern Europe, and the Mediterranean appears to be more closely represented by the cVAE, compared to SEAS5. This aligns with positive skill values when referenced to the climatology and SEAS5 in the Mediterranean, especially the eastern part. On the contrary, although a better correlation pattern is found in the Nordic countries, this does not directly translate into skill improvements in our assessment.

Discussion

This study proposed and evaluated a novel data-driven method for predicting seasonal climate anomalies. Our methodology combines variational inference with vision transformers to generate ensembles of global seasonal predictions (lead months 1-3). Global skill scores show that our temperature predictions surpass the predictability provided by climate forcing trends and even outperform SEAS5 in some ocean and land areas, of which the latter are particularly relevant for potential user applications. Precipitation forecasts exhibit a more limited skill compared to climatology, and SEAS5 clearly outperforms our data-driven approach in the equatorial band. Still, we observe similar spatial patterns of skill for both temperature and precipitation compared to SEAS5, suggesting that both systems build on similar sources of predictability. The model's predictive skill beyond trend-based forecasts further validates our methodological approach, trained on (imperfect) CMIP6 simulations, and its capability in simulating, to some extent, interannual variability.

We also highlight the advantages of training the prediction model on smaller regions, enabling a more tailored optimisation of regional-specific features. Our case study focuses on Europe, an area that often struggles with accurate numerical seasonal predictions and demonstrates improvements over a global model. Our methodology outperforms SEAS5 in predicting European temperatures in multiple sub-regions and seasons. However, precipitation forecasts from both dynamical and data-driven prediction systems exhibit limited skill. Overall, we find that increasing spatial resolution or constraining the target region provides benefits without compromising prediction quality, enabling flexible configuration choices based on user needs and computational constraints. Furthermore, predictions of the El Niño-Southern Oscillation (ENSO) and the North Atlantic Oscillation (NAO) indices show comparable skill and variability to SEAS5, and correlation maps between both indices and their respective predicted temperature and rainfall fields align largely with those found in SEAS5 and ERA5. The consistency of both the regional impacts of the teleconnections and the spatial patterns of skill across different target configurations reinforces the robustness of our methodology, suggesting that these patterns represent genuine features in the climate system rather than artefacts of the machine learning model.

Previous research[38] has shown that similar variational architectures can provide skilful seasonal predictions for the October to March seasonal mean (lead months 2-7). As this previous research pointed out, predicting a longer seasonal average forecast is relatively straightforward compared to our setup. Part of this is due to longer averages filtering out higher-frequency climate variability, and the more distant lead time reducing the influence of the initialisation, making initialised dynamical prediction systems a weaker baseline[11,39]. Besides, they highlighted a pitfall in their strategy of splitting the training and validation datasets, where they apply random shuffling through the entire set of CMIP simulations. This splitting strategy is prone to introducing autocorrelations between the training and validation sets due to the persistent impact of low-frequency climate signals. By addressing a more difficult prediction task, three-month averages (lead months 1-3), properly splitting distinct simulations in our train and validation split, accounting for long-term forcing trends in our verification, and studying fundamental modes of variability and teleconnections, we aim to increase trust in this and similar methodologies

As an example of an alternative data-driven approach, a recent study[55] published during the review of this manuscript has leveraged ACE2[56] to produce seasonal predictions (also lead months 1-3, but for DJF). While our approach trains a model on climate model output to predict the 3-month average anomaly in a single step, this alternative setup builds on the success of AI-based weather forecasts trained on ERA5 to predict the evolution of the atmosphere at 6-hour intervals, thereby remaining stable over long forecast periods. In terms of skill, that study reached similar conclusions to the ones of this manuscript - ACE2 shows slightly lower but comparable skill to GloSea[57] (GC3.2 configuration), another state-of-the-art dynamical prediction system. However, the contribution of trends to the skill obtained is not assessed. ACE2's strength stems from its training on ERA5 reanalysis, which helps avoid model errors and the misrepresentation of certain teleconnections, both inherent to climate simulations. However, this advantage comes at the cost of retaining only 10 test samples (2001-2010) outside the training data. Thus, the combination of a small test set and a training set drawn from years after the testing period (from 2010 onwards) raises concerns about the model's ability to extrapolate to unseen scenarios and the reliability of its verification[51].

Likewise, important limitations must be acknowledged in our work. Data-driven approaches that rely on climate model output during training are susceptible to learning biases or model errors from it, limiting their performance. Combining outputs from different dynamical models can help compensate some of those model-specific errors. Yet, errors that are systematic across models will still be learned by our approach and similar ones. Additionally, our method employs a minimal initialisation and output setup, comprising of monthly and seasonal averages, as well as a limited variable set. This setup is far from current operational prediction systems based on dynamical climate models, which utilise an extensive set of 3D atmospheric, land and ocean state variables for their initialisation. Such differences in the initialisation process could contribute to SEAS5's higher skill compared to our data-driven approach in some regions/features. In addition, although the teleconnection assessment presented in this manuscript helps explain some of the skill patterns observed in our predictions, we acknowledge the limitations of this initial examination and call for follow-up experiments to understand the learned relationships better and evaluate spatial-temporal relationships and causality in data-driven climate predictions.

Some of these limitations represent opportunities to enhance our approach. For example, errors in climate model output can be mitigated through fine-tuning or guidance techniques[58], performance sub-selection of the different simulations used during the training stage, or by incorporating improved simulations, such as those from novel high-resolution climate models[59]. In this latter case, generative models could be especially valuable for saving substantial computational resources or even enabling the operationalisation of such predictions. Score-based[60] (diffusion)[61] or flow-matching[62] approaches can provide better modelling of the conditional probabilities predicted, potentially addressing some of the optimisation challenges inherent in variational inference. Additionally, we acknowledge that the presented methodology constitutes a prototype, and better initialisation with multiple variables at a higher temporal frequency is perfectly implementable and could further improve our prediction system, bringing it closer to state-of-the-art dynamical operational prediction systems.

Thus, this study advances our initial objective of further developing probabilistic deep learning methods for seasonal prediction, demonstrating that generative models trained on climate models can achieve comparable skill to current operational dynamical prediction systems. While

challenges remain to further enhance the performance and possibly outperform current state-of-the-art prediction systems, our results establish a promising foundation for the future development of data-driven and seasonal prediction systems.

Methods

Problem formulation

The objective is to predict the climate state $Y \in R^{c_y \times n_{lat} \times n_{lon}}$ of a future season based on current and past states $X^i \in R^{c \times n_{lat} \times n_{lon}}$ from the i preceding months. To deal with the stochastic nature of the atmosphere beyond 12 days[1] we intend to forecast not a deterministic value but the conditional probability distribution $p(Y^{t+1}|X^t,X^{t-1},\ldots,X^{t-T})$ of the target season Y^{t+1} on the current and previous conditions X.

The representation of the target season Y is comprised by c_y variables of 3-month seasonal averages on a 5°×5° latitude-longitude grid. As stated in[41], this grid-scale is a good compromise between capturing the large-scale climate signal and smoothing out noise while saving computational resources. The representation of the initial states X is comprised by c_x variables of monthly averages on a 5°×5° latitude-longitude grid. The same methodology can be tested under different representations, combining different grids and temporal resolutions (both at source and output) due to its inherent flexibility. Thus, for the regional use case, we increase the spatial resolution of the target Y to 1°.

Variational Inference

To obtain the conditional probability distribution $p(Y|X)$ on a target climate state Y given a state X of current or past conditions, state-of-the-art climate prediction systems run multiple dynamical simulations, each with slightly perturbed initial conditions, obtaining an ensemble of plausible outcomes from which probabilities can be inferred. Analogously, our objective is to learn a statistical model $p_\theta(y|x,z)$ from which multiple predictions can be inferred statistically from a set

of initial conditions and an n-dimensional latent variable z that adds the stochastic component to the statistical model.

Learning the conditional probability distribution $p_\theta(y|x)$ from data is not a straightforward problem[63]. Ideally, we would like to minimise the difference between our learned distribution $p_\theta(y|x)$ and the observed data distribution $q_D(y|x)$. This objective can be achieved empirically by maximising the sum over the log-likelihoods of our data points in the learned distribution[64]. Yet, this is computationally intractable as it requires integration over z for each data point:

$$p_\theta(y|x) = \int p_\theta(y|z,x)p_\theta(z|x)dz \qquad (1)$$

Amortised variational inference[43] offers an alternative by narrowing the integration space of z to values that are likely to generate y. This likelihood is described by $p(z|y_i,x_i)$ and is approximated by an amortised inference distribution $q_\phi(z|y,)$ that is also learned. To jointly optimise the parameters ϕ and θ a lower-bound of the log-likelihood or evidence lower-bound (ELBO) is defined:

$$L(\theta,\varphi) = -\mathbb{E}_{q_\varphi(z|x,y)}[\log p_\theta(y|z,x)] + D_{KL}\big(q_\varphi(z|y,x) \,||\, p_\theta(z|x)\big) \qquad (2)$$

where $E_{q_\phi(z|x,y)}[\log p_\theta(y|z,x)]$ is the expected log-likelihood of y given z and x, and $D_{KL}(q_\phi(z|y,x)||p_\theta(z|x))$ is the KL divergence between the approximate posterior $q_\phi(z|y,x)$ and the prior $p_\theta(z|x)$.

Thus, our final objective is to jointly train two neural networks: $q_\phi(z|y,x)$ representing the learned approximate posterior, and $p_\theta(y|z,x)$ being the learned generative model. $q_\phi(z|y,x)$ will be represented by an encoder applied to the target state Y and only used during the training phase. While $p_\theta(y|z,x)$ will be conformed by an encoder on the initial state X and the decoder generating new predictions Y_θ combining information from the learned latent space and the compact

representation of the initial state χ. Minimising this ELBO allows joint optimisation of θ and ϕ, effectively approximating the intractable $p_\theta(\mathcal{Y}|\mathcal{X})$.

Architecture

The model architecture design is essential for extracting meaningful features that improve seasonal predictability. The architecture needs to capture both temporal and spatial long-range interactions influenced by global teleconnections, as well as local interactions that stem from land-atmosphere processes and persistence. However, due to the limited size of the available training dataset, keeping the model complexity in check is essential to avoid overfitting. Choosing the architectural design implies finding a sustainable balance between these factors.

Vision Transformers (ViTs) are a well-suited option for this task[45,65]. ViTs employ a general-purpose inductive bias that allows them to model distant and local connections without needing the deep hierarchy and pooling operations typical of Convolutional Neural Networks (CNNs). Thus, they decouple the interaction range from the network depth, and this is particularly helpful when modelling the different types of interactions that occur at seasonal time scales. In addition, transformers are very suitable for incorporating data with different formats (i.e. time series with 2D or 3D spatial grids). They can also make inferences even under the erratic presence of missing values. Still, due to their unconstrained non-locality, ViTs are known to need large datasets in order to train correctly. These reasons partly explain the multiple applications of ViTs found in weather prediction[66–68], contrasting the few to no applications for seasonal prediction.

As illustrated in Figure 1 panel B, our model architecture combines the variational inference framework of a conditional Variational Autoencoder (cVAE) with ViT encoders for feature extraction. The $q_\phi(z|\mathcal{Y},\mathcal{X})$ approximate posterior is represented by a ViT encoder applied to the target climate state γ. This encoder generates a compact latent representation z_γ, which is then passed through a Multi-Layer Perceptron (MLP) to produce the variational parameters μ and σ that will conform the learned posterior distribution $q_\phi(z|\mathcal{Y},\mathcal{X})$. This network component (depicted in orange) is only used during training. The $p_\theta(\mathcal{Y}|z,\mathcal{X})$ generative model is formed by an additional

ViT encoder applied to the initial climate state χ, producing a reduced representation z_x. This reduced representation z_x latent is then combined with a sample from the posterior distribution $q_\phi(z|y,x)$ and passed through a Convolutional Neural Network (CNN) decoder to generate new climate predictions γ_θ.

Once the model is trained, deterministic predictions can be generated by sampling values from the prior distribution z. Each sample z is concatenated with the z_x latent representing the initial state and decoded through the CNN decoder, obtaining a deterministic prediction (or ensemble member) conditioned on the initial state x. By repeating this process with multiple samples of z, we draw the $p_\theta(y|z,x)$ distribution learned by the model, obtaining an ensemble of predictions that capture the uncertainty of the system. This architecture allows the model to extract meaningful features from the input data while maintaining a constrained overall size. By jointly optimizing the encoder (q_ϕ) and decoder (p_θ) networks using minimizing loss objective, the model learns to generate diverse, physically consistent ensemble predictions while capturing the underlying uncertainty in the data.

Datasets

We use four different climate models (see Table 1) from the Coupled Model Intercomparison Project 6 (CMIP6[47]) to obtain a sufficiently large training set. The historical and SSP2-4.5 scenarios are concatenated for each realisation into a continuous time series spanning 1880 to 2080. These specific models were chosen as they meet the criteria for the number of realisations and output variables. All the models' output was obtained from the Earth System Grid Federation (ESGF) and gathered and pre-processed to joint spatial resolution and units using ESMValtool[69].

Split	Source	Time Period	Models
Training	CMIP6 (Hist. + SSP245)	1880 - 2080	CanESM5 r(6:25)i1p1f1, CanESM5 r(6:25)i1p2f1, MIROC-ES2L _r(6:25)i1p1f2, MIROC6 r(6:25)i1p1f1 & MPI-ESM1-2-LR r(6:25)i1p1f1

Validation	CMIP6 (Hist. + SSP245)	1880 - 2080	CanESM5 r(1:5)i1p1f1, CanESM5 r(1:5)i1p2f1, MIROC-ES2L _r(1:5)i1p1f2, MIROC6 _r(1:5)i1p1f1 & MPI-ESM1-2-LR r(1:5)i1p1f1
Test	ERA5	1950 - 2021	

Table 1. Datasets description information

For the evaluation of the data-driven models, we use the ERA5[48] reanalysis covering 1950 to 2021. ERA5 is produced using 4D-Var data assimilation combined with the ECMWF Integrated Forecast System (IFS) CY41R2. Again, ERA5 data was pre-processed to a common spatial grid and units using ESMValtool.

We also use the ECMWF's seasonal climate prediction SEAS5[10] as a dynamical benchmark against the data-driven models. SEAS5 is based on the Integrated Forecast System (IFS) atmospheric component coupled to the Nucleus for European Modelling of the Ocean (NEMO) ocean model and the dynamic Louvain-la-Neuve Sea Ice Model (LIM2). SEAS5 operational seasonal forecasts are initialised on the first day of each month, and 51 ensembles are initialised covering up to seven months in the future. Additionally, a set of hindcasts (1981 to 2017) are also produced with the same configuration but with a reduced ensemble (25 realisations). As a benchmark, we concatenate both hindcast and forecasts into a continuous set of forecasts covering the period 1981 to 2021 with an ensemble of 25 realisations. The SEAS5 data was obtained from the Copernicus Climate Data Store (CDS) API.

Data preprocessing

We perform an initial homogenization of all the climate model output and reanalysis data to a common spatio-temporal resolution and units. For both inputs and outputs, monthly means at 5° or 1° horizontal resolution are used (depending on the prediction task).

The second stage involves the standardisation of the data, including de-trending, anomaly computation, and normalisation. Data standardisation is a critical aspect of data-driven climate predictions. On one hand, the standardisation of inputs and outputs can drastically change the prediction task assigned to the model (i.e. predicting seasonal averages over the trend vs

predicting the forcing influence at the seasonal time scale). Similarly, it can affect the models' performance, as the standardisation can remove or add information from the climate fields used. Finally, improper standardisation of the outputs, references, and benchmarks can lead to misleading claims of performance during the validation[42,51], especially under strong trends[41,70]. At the same time, data standardisation helps in the speed and stability during the training of data-driven approaches.

De-trending for CMIP6 data is performed removing the forced component (ensemble mean) of each model independently. For the ERA5 reanalysis, locally estimated scatterplot smoothing (LOESS)[71], with a fixed time window of 30 years and one degree of freedom, is applied to obtain a non-linear trend later removed from the data. To avoid overestimates of forecast skill due to the use of information not available at forecast time[51], we fit LOESS using only values prior to the forecasting time (retroactively). As shown in Figure 1, de-trending is applied to the target Y during training, as well as during the validation of the forecasts.

Standardised monthly anomalies are computed by subtracting the mean and normalising by the standard deviation of the 1981 to 2000 period. As an exemption, precipitation values are fitted to a gamma distribution instead. All these steps are applied point by point to each climate model and reanalysis independently, helping to remove significant biases present in both climate models and the reanalysis output. However, none of the data from the testing period 2001-2021 is included in this process.

Assessing forecasts quality

In this work, we use a set of verification metrics to quantify the quality of the predictions developed and we compare the results against the ECMWF state-of-the-art seasonal prediction system SEAS5. As exposed in [41], we identify two main objectives. First, to assess whether our proposed model produces more accurate predictions compared to a reference forecast, in our case a state-of-the-art dynamical forecasting system. Second, to assess whether the ensemble spread of our method provides a good estimation of uncertainty on average.

The first objective can be fulfilled by employing deterministic metrics. As a first individual assessment of the different forecasts, we employ the Spearman correlation (Equation 1) between the anomalies of the ensemble median of our predictions and the ground truth, also known as the Anomaly Correlation Coefficient (ACC). It helps quantify the monotonic relationship between these two. The Spearman correlation is preferred over the Pearson correlation due to its non-parametric nature and insensitivity to outliers.

$$\rho = 1 - \frac{6\sum_{i=1}^{n}\left(r_{x_i} - r_{y_i}\right)^2}{n(n^2-1)} \quad (3)$$

where r_x is the ranks of the predictions' ensemble median, r_y the ranks of the ground truth, and n the number of samples.

In addition, the root mean square error (RMSE) is used to add information of the potential mean and conditional biases in our predictions (Equation 2). The RMSE can be expressed as a function of the Spearman correlation and the mean and conditional biases[72], providing a complete deterministic overview of our predictions.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i^2 - y_i^2)} \quad (4)$$

where $\hat{x}$ is the ensemble median of our predictions, y are the observations, and n the number of samples.

Our second objective is better fulfilled using probabilistic metrics, which test whether the spread in our prediction is adequate to quantitatively represent the range of possibilities for individual predictions over time. We base our probabilistic validation on the Continuous Ranked Probability (CRPS), a measure of squared error in the probability space:

$$\text{CRPS}(P, y) = \int_{-\infty}^{\infty}\left[F(x) - H_y(x)\right]^2 dx \quad (5)$$

where F is the proposed cumulative distribution function (CDF) obtained from the forecast ensemble x, and H *is* the Heaviside step function centred at the actual observed value y.

To facilitate the interpretability and comparison of the results, both CRPS and RMSE are expressed as skill scores referenced against a climatological forecast (clim) or SEAS5:

$$CRPSS = 1 - \frac{CRPS}{CPRS_{ref}} \quad RMSS = 1 - \frac{RMSE}{RMSE_{ref}} \qquad (6,7)$$

Uncertainty in the validation metrics is evaluated using a non-parametric bootstrapping approach. The forecasts and reanalysis observations are reshuffled in this method to compute 1000 core values. For the ACC computation, the values obtained are compared to the 95th significance level against a similar distribution generated using a random time series instead of the forecast. For the skill score metrics, we reshuffled the forecast, reference forecast, and reanalysis time series to compute a distribution of skill scores. Then, we assess whether the score value is significantly greater than zero at the 95th significance level.

Architecture & training configuration

The model processes five key climate variables: 2-meter air temperature (tas), precipitation (pr), sea surface temperature (tos), and geopotential height at 500hPa and 300hPa levels (zg500, zg300). For the input state X, these variables represent conditions during the preceding 6 months, while the target state Y is comprised of the same variables seasonal averaged comprised by lead months 1 to 3. Topography, land-ocean, and encoded latitude and longitude coordinates are also concatenated to the input state X. During inference, 150 ensembles are pooled by sampling from the latent space and conditioned on the inputs.

Our conditional Variational Autoencoder (cVAE) implements dual Vision Transformer (ViT) encoders to process input and target climate states. Each encoder pathway consists of 8 transformer layers with an embedding dimension of 128 and single-head attention, operating on patch sizes of 1 to capture fine-grained spatial features. The latent space has a dimension of 128,

enabling a compact representation of climate patterns. The decoder employs four residual blocks with 32 filters each, using convolutional layers to reconstruct the predicted climate fields.

We train a separate model for each season, concatenating adjacent seasons (±1 month) to increase the training set size. Each model underwent training for 50 epochs with a batch size of 256, using an initial learning rate of $1e^{-4}$ and weight decay of 0.001 for regularization. Models are optimized through an information maximization loss function for variational autoencoders, or InfoVAE objective[73], that constitutes a generalization of the ELBO objective. The reconstruction term is weighted by the cosine of the latitude to account for differences in grid-cell area. The weighting parameter (lambda) was set to 1 and the confidence parameter (alpha) to 0.9. This configuration was selected according to the results obtained in the validation set when sampling pseudo-randomly different hyper-parameter configurations (not shown).

The basic configuration and hyper-parameters of the model are consistent across the experiments shown in panel D of Figure 1, with one exception: in the Regional 5° model, the number of blocks in the decoder is reduced from 4 to 2 due to the limited number of points. The training stage of the global 5° configuration takes 3.3 hours using 4xH100 NVIDIA GPUs. Training times of similar magnitude are obtained for the other target configurations. Inferencing a complete hindcast (1950-2021) for all variables and a single season takes seconds on a single H100 GPU across all model configurations. Additional details regarding the architecture and hyper-parameters can be found in the Supplementary Information (Table S1).

Data availability

All the data used are publicly available or restricted to the signed-up users. SEAS5 and ERA5 data were downloaded from the official website of Copernicus Climate Data (CDS) at https://cds.climate.copernicus.eu/. CMIP6 datasets were downloaded from the Earth System Grid Federation (ESGF). CMIP6 and ERA5 datasets were pre-processed using ESMValTool (https://esmvaltool.org).

Code availability

The code used for data processing, model training, inference, and evaluation is available at https://gitlab.earth.bsc.es/es/seasgen/.

Acknowledgements

This work was supported by the AI4Drought (ESA AI4SCIENCE; contract number 4000137110/22/I-EF) and CERISE (European Union; grant agreement No101082139). AD holds a fellowship within the "Generación D" initiative, Red.es, Ministerio para la Transformación Digital y de la Función Pública, for talent attraction (C005/24-ED CV1), funded by the European Union NextGenerationEU funds, through PRTR. MGD and SM are grateful for support from the Horizon Europe project EXPECT (Grant 101137656). The authors thank Pierre-Antoine Bretonnier and Margarida Samsó for their assistance in downloading and formatting part of the data.

Author contributions

Ll.P., A.D. and M.D. conceived the research idea. Ll.P. and A.P. implemented the deep learning code. Ll.P. Implemented the validation pipeline. D.C. contributed to the development of the code. Ll.P. drafted the manuscript with input from all co-authors. All authors discussed the results and revised the manuscript. M.D., S.M., A.M., J.P., L.R. and A.S. supervised the project.

Competing Interests

The authors declare no competing interests.

References

1. Lorenz, E. N. Deterministic Nonperiodic Flow. https://journals.ametsoc.org/view/journals/atsc/20/2/1520-0469_1963_020_0130_dnf_2_0_co_2.xml (1963).
2. Palmer, T. N. & Anderson, D. L. T. The prospects for seasonal forecasting—A review paper. Q. J. R. Meteorol. Soc. 120, 755–793 (1994).

3. Challinor, A. J., Slingo, J. M., Wheeler, T. R. & Doblas-Reyes, F. J. Probabilistic simulations of crop yield over western India using the DEMETER seasonal hindcast ensembles. Tellus Dyn. Meteorol. Oceanogr. 57, 498–512 (2005).
4. Pérez-Zanón, N. et al. Lessons learned from the co-development of operational climate forecast services for vineyards management. Clim. Serv. 36, 100513 (2024).
5. García-Morales, M. B. & Dubus, L. Forecasting precipitation for hydroelectric power management: how to exploit GCM's seasonal ensemble forecasts. Int. J. Climatol. 27, 1691–1705 (2007).
6. Lledó, L., Cionni, I., Torralba, V., Bretonnière, P.-A. & Samsó, M. Seasonal prediction of Euro-Atlantic teleconnections from multiple systems. Environ. Res. Lett. 15, 074009 (2020).
7. Ramon, J., Lledó, L., Bretonnière, P.-A., Samsó, M. & Doblas-Reyes, F. J. A perfect prognosis downscaling methodology for seasonal prediction of local-scale wind speeds. Environ. Res. Lett. https://doi.org/10.1088/1748-9326/abe491 (2021) doi:10.1088/1748-9326/abe491.
8. Thomson, M. C. et al. Malaria early warnings based on seasonal climate forecasts from multi-model ensembles. Nature 439, 576–579 (2006).
9. Saha, S. et al. The NCEP Climate Forecast System Version 2. J Clim 27, 2185–2208 (2014).
10. Johnson, S. J. et al. SEAS5: The new ECMWF seasonal forecast system. Geosci. Model Dev. 12, 1087–1117 (2019).
11. Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P. & Rodrigues, L. R. L. Seasonal climate predictability and forecasting: Status and prospects. Wiley Interdiscip. Rev. Clim. Change 4, 245–268 (2013).
12. Meehl, G. A. et al. Initialized Earth System prediction from subseasonal to decadal timescales. Nat. Rev. Earth Environ. 2, 340–357 (2021).
13. Scaife, A. A. et al. Does increased atmospheric resolution improve seasonal climate predictions? Atmospheric Sci. Lett. 20, e922 (2019).
14. Materia, S. et al. Impact of Atmosphere and Land Surface Initial Conditions on Seasonal Forecasts of Global Surface Temperature. https://doi.org/10.1175/JCLI-D-14-00163.1 (2014) doi:10.1175/JCLI-D-14-00163.1.
15. Ardilouze, C. et al. Multi-model assessment of the impact of soil moisture initialization on mid-latitude summer predictability. Clim. Dyn. 49, 3959–3974 (2017).
16. Weisheimer, A. & Palmer, T. N. On the reliability of seasonal climate forecasts. J. R. Soc. Interface 11, 20131162 (2014).
17. Eden, J. M., van Oldenborgh, G. J., Hawkins, E. & Suckling, E. B. A global empirical system for probabilistic seasonal climate prediction. Geosci. Model Dev. 8, 3947–3973 (2015).
18. Hao, Z., Singh, V. P. & Xia, Y. Seasonal Drought Prediction: Advances, Challenges, and Future Prospects. Rev. Geophys. 56, 108–141 (2018).
19. Andersson, T. R. et al. Seasonal Arctic sea ice forecasting with probabilistic deep learning. Nat. Commun. 2021 121 12, 1–12 (2021).
20. Gibson, P. B. et al. Training machine learning models on climate model output yields skillful interpretable seasonal precipitation forecasts. Commun. Earth Environ. 2, 159 (2021).

21. Materia, S. et al. Artificial intelligence for climate prediction of extremes: State of the art, challenges, and future perspectives. WIREs Clim. Change e914 (2024) doi:10.1002/wcc.914.
22. Bi, K. et al. Pangu-Weather: A 3D High-Resolution Model for Fast and Accurate Global Weather Forecast. Preprint at https://doi.org/10.48550/arXiv.2211.02556 (2022).
23. Pathak, J. et al. FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators. Preprint at https://doi.org/10.48550/arXiv.2202.11214 (2022).
24. Lam, R. et al. Learning skillful medium-range global weather forecasting. Science 382, 1416–1421 (2023).
25. Price, I. et al. GenCast: Diffusion-based ensemble forecasting for medium-range weather. Preprint at https://doi.org/10.48550/arXiv.2312.15796 (2024).
26. Kochkov, D. et al. Neural general circulation models for weather and climate. Nature 632, 1060–1066 (2024).
27. Ham, Y.-G., Kim, J.-H. & Luo, J.-J. Deep learning for multi-year ENSO forecasts. Nature 573, 568–572 (2019).
28. Felsche, E. & Ludwig, R. Applying machine learning for drought prediction in a perfect model framework using data from a large ensemble of climate simulations. Nat. Hazards Earth Syst. Sci. 21, 3679–3691 (2021).
29. Ding, H., Newman, M., Alexander, M. A. & Wittenberg, A. T. Skillful Climate Forecasts of the Tropical Indo-Pacific Ocean Using Model-Analogs. https://doi.org/10.1175/JCLI-D-17-0661.1 (2018) doi:10.1175/JCLI-D-17-0661.1.
30. Mahmood, R. et al. Constraining low-frequency variability in climate projections to predict climate on decadal to multi-decadal timescales – a poor man's initialized prediction system. Earth Syst. Dyn. 13, 1437–1450 (2022).
31. Donat, M. G., Mahmood, R., Cos, P., Ortega, P. & Doblas-Reyes, F. Improving the forecast quality of near-term climate projections by constraining internal variability based on decadal predictions and observations. Environ. Res. Clim. 3, 035013 (2024).
32. Cos, P., Marcos-Matamoros, R., Donat, M., Mahmood, R. & Doblas-Reyes, F. J. Near-Term Mediterranean Summer Temperature Climate Projections: A Comparison of Constraining Methods. J. Clim. 37, 4367–4388 (2024).
33. Rader, J. K. & Barnes, E. A. Optimizing Seasonal-To-Decadal Analog Forecasts With a Learned Spatially-Weighted Mask. Geophys. Res. Lett. 50, e2023GL104983 (2023).
34. Belkin, M., Hsu, D., Ma, S. & Mandal, S. Reconciling modern machine learning practice and the bias-variance trade-off. Proc. Natl. Acad. Sci. 116, 15849–15854 (2019).
35. Curth, A. Classical Statistical (In-Sample) Intuitions Don't Generalize Well: A Note on Bias-Variance Tradeoffs, Overfitting and Moving from Fixed to Random Designs. Preprint at http://arxiv.org/abs/2409.18842 (2024).
36. Ling, F. et al. Multi-task machine learning improves multi-seasonal prediction of the Indian Ocean Dipole. Nat. Commun. 2022 131 13, 1–9 (2022).
37. Beobide-Arsuaga, G., Düsterhus, A., Müller, W. A., Barnes, E. A. & Baehr, J. Spring Regional Sea Surface Temperatures as a Precursor of European Summer Heatwaves. Geophys. Res. Lett. 50, e2022GL100727 (2023).

38. Pan, B. et al. Improving Seasonal Forecast Using Probabilistic Deep Learning. J. Adv. Model. Earth Syst. 14, (2022).
39. Patterson, M., Weisheimer, A., Befort, D. J. & O'Reilly, C. H. The strong role of external forcing in seasonal forecasts of European summer temperature. Environ. Res. Lett. 17, 104033 (2022).
40. Tippett, M. K. & Becker, E. J. Trends, Skill, and Sources of Skill in Initialized Climate Forecasts of Global Mean Temperature. Geophys. Res. Lett. 51, e2024GL110703 (2024).
41. Goddard, L. et al. A verification framework for interannual-to-decadal predictions experiments. Clim. Dyn. 40, 245–272 (2013).
42. Risbey, J. S. et al. Common Issues in Verification of Climate Forecasts and Projections. Climate 10, 83 (2022).
43. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. Preprint at https://doi.org/10.48550/ARXIV.1312.6114 (2013).
44. Kingma, D. P. & Welling, M. An Introduction to Variational Autoencoders. https://doi.org/10.48550/ARXIV.1906.02691 (2019) doi:10.48550/ARXIV.1906.02691.
45. Dosovitskiy, A. et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Preprint at https://doi.org/10.48550/arXiv.2010.11929 (2021).
46. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature 521, 436–444 (2015).
47. Eyring, V. et al. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. Geosci. Model Dev. 9, 1937–1958 (2016).
48. Hersbach, H. et al. The ERA5 global reanalysis. Q. J. R. Meteorol. Soc. 146, 1999–2049 (2020).
49. Kim, H.-M., Webster, P. J. & Curry, J. A. Seasonal prediction skill of ECMWF System 4 and NCEP CFSv2 retrospective forecast for the Northern Hemisphere Winter. Clim. Dyn. 39, 2957–2973 (2012).
50. Kumar, A., Chen, M. & Wang, W. Understanding Prediction Skill of Seasonal Mean Precipitation over the Tropics. https://doi.org/10.1175/JCLI-D-12-00731.1 (2013) doi:10.1175/JCLI-D-12-00731.1.
51. Risbey, J. S. et al. Standard assessments of climate forecast skill can be misleading. Nat. Commun. 12, (2021).
52. Fisher, R. A. Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. Biometrika 10, 507–521 (1915).
53. Iturbide, M. et al. An update of IPCC climate reference regions for subcontinental analysis of climate model data: definition and aggregated datasets. Earth Syst. Sci. Data 12, 2959–2970 (2020).
54. Yang, Z. & Villarini, G. Examining the capability of reanalyses in capturing the temporal clustering of heavy precipitation across Europe. Clim. Dyn. 53, 1845–1857 (2019).
55. Kent, C. et al. Skilful global seasonal predictions from a machine learning weather model trained on reanalysis data. Npj Clim. Atmospheric Sci. 8, 314 (2025).
56. Watt-Meyer, O. et al. ACE2: accurately learning subseasonal to decadal atmospheric variability and forced responses. Npj Clim. Atmospheric Sci. 8, 205 (2025).
57. Global Seasonal forecast system version 5 (GloSea5): a high-resolution seasonal forecast system - MacLachlan - 2015 - Quarterly Journal of the Royal Meteorological Society - Wiley Online Library. https://rmets.onlinelibrary.wiley.com/doi/full/10.1002/qj.2396.

58. Kolloff, C. et al. Minimum-Excess-Work Guidance. Preprint at https://doi.org/10.48550/arXiv.2505.13375 (2025).
59. Rackow, T. et al. Multi-year simulations at kilometre scale with the Integrated Forecasting System coupled to FESOM2.5 and NEMOv3.4. Geosci. Model Dev. 18, 33–69 (2025).
60. Song, Y. et al. Score-Based Generative Modeling through Stochastic Differential Equations. Preprint at https://doi.org/10.48550/arXiv.2011.13456 (2021).
61. Ho, J., Jain, A. & Abbeel, P. Denoising Diffusion Probabilistic Models. Preprint at https://doi.org/10.48550/arXiv.2006.11239 (2020).
62. Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M. & Le, M. Flow Matching for Generative Modeling. Preprint at https://doi.org/10.48550/arXiv.2210.02747 (2023).
63. Murphy, K. P. Probabilistic Machine Learning : Advanced Topics.
64. Wilks, D. S. Statistical Methods in the Atmospheric Sciences: An Introduction. (Elsevier, Amsterdam, 2019).
65. Vaswani, A. et al. Attention Is All You Need. Preprint at https://doi.org/10.48550/arXiv.1706.03762 (2023).
66. Bodnar, C. et al. Aurora: A Foundation Model of the Atmosphere. Preprint at http://arxiv.org/abs/2405.13063 (2024).
67. Nguyen, T. et al. Scaling transformer neural networks for skillful and reliable medium-range weather forecasting. Preprint at https://doi.org/10.48550/arXiv.2312.03876 (2023).
68. Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K. & Grover, A. ClimaX: A foundation model for weather and climate. Preprint at http://arxiv.org/abs/2301.10343 (2023).
69. Righi, M. et al. Earth System Model Evaluation Tool (ESMValTool) v2.0 – technical overview. Geosci. Model Dev. 13, 1179–1199 (2020).
70. Wulff, C. O., Vitart, F. & Domeisen, D. I. V. Influence of trends on subseasonal temperature prediction skill. Q. J. R. Meteorol. Soc. 148, 1280–1299 (2022).
71. Mahlstein, I., Spirig, C., Liniger, M. A. & Appenzeller, C. Estimating daily climatologies for climate indices derived from climate model data and observations. https://doi.org/10.1002/2014JD022327 doi:10.1002/2014JD022327.
72. Murphy, Alan. Skill scores based on the mean square error and their relationships to the correlation coefficient. Mon. Weather Rev. 116, (1988).
73. Zhao, S., Song, J. & Ermon, S. InfoVAE: Information Maximizing Variational Autoencoders. Preprint at https://doi.org/10.48550/arXiv.1706.02262 (2018).

Fig. 1. Methodology overview. A, Illustration of the signal decomposition of the target variable Y. B, Schematic representation of the conditional Variational Autoencoder (cVAE) architecture. Two vision transformers (ViTs) encode the information from multiple climate fields into the latent space. The compressed latent space representation is then passed to the CNN decoder that reconstructs the predicted climate fields. C, Final

model assembling, combining the interannual variability prediction from the cVAE model and the regressed LOESS trend. D, tested model configurations, combining different spatial resolutions and target domains.

Fig. 2. ACC against ERA5 reanalysis. ACC against ERA5 reanalysis (2001–2021), for temperature (2T, top row) and precipitation rate (PR, bottom row). Seasons are shown in the columns: DJF (December–January–February), MAM (March–April–May), JJA (June–July–August), and SON (September–October–November). Black dots indicate statistical significance at the 95% confidence level.

Fig. 3. Ensemble size effect on model's deterministic performance. Effect of ensemble size on the ACC of the ensemble median (blue thick line) against ERA5 (red thick line). The figure depicts MAM temperature anomalies in a grid cell located at 42.5°N, 10.5°E. Blue thinner lines depict the individual ensembles.

Fig. 4. Global temperature skill scores. Forecast skill scores for near-surface air temperature (tas) predictions from 2001–2021, using 1981–2000 as the reference period of the climatological forecast and anomalies. The panels show four seasons (DJF, MAM, JJA, SON) across columns and two skill metrics: the root-mean-square error skill score (RMSS) and the continuous ranked probability skill score (CRPSS). Skill scores range from 0 (pink, indicating no skill) to 0.5 (dark blue, indicating high skill above the reference). Metrics are referenced (Ref.) against the climatological forecasts (CLIM) or against the ECMWF's seasonal prediction system (SEAS5). Black dots indicate statistical significance of the skill score being positive at the 95% confidence level (more details can be found in the Methods section).

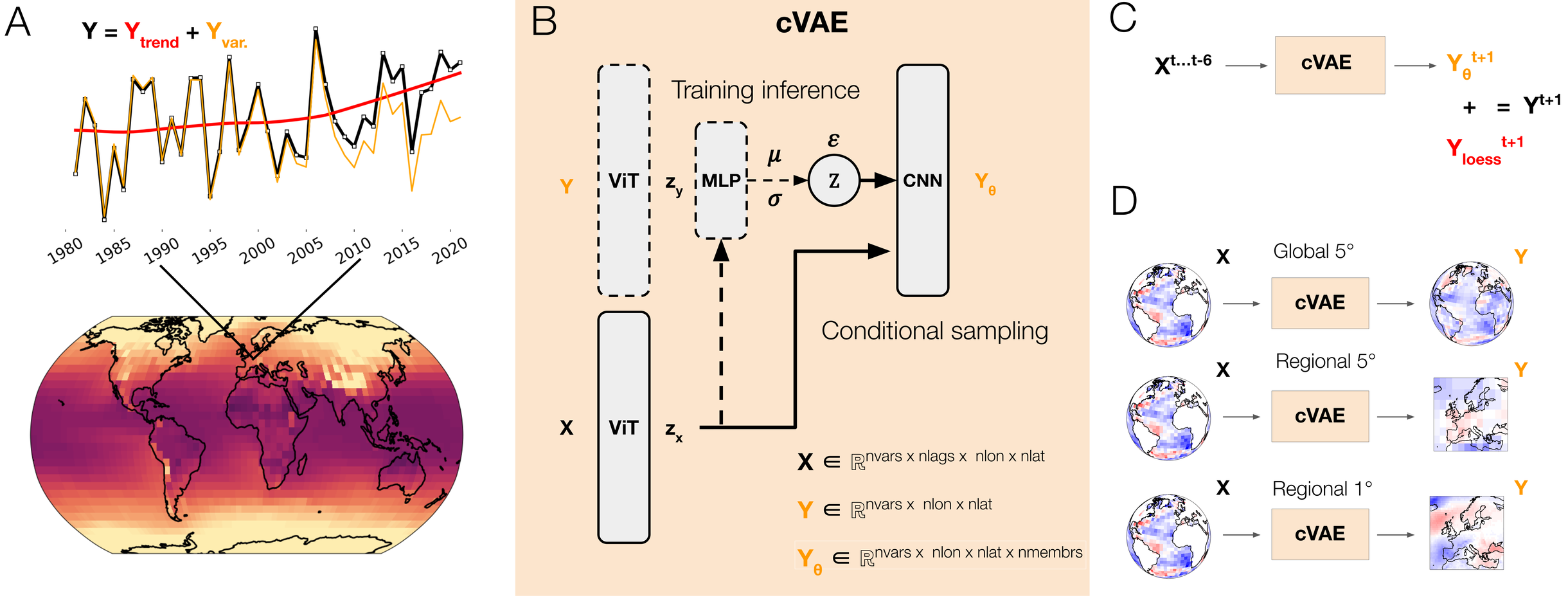
Fig. 5. Global precipitation skill scores. Same as Figure 4 but for precipitation predictions.

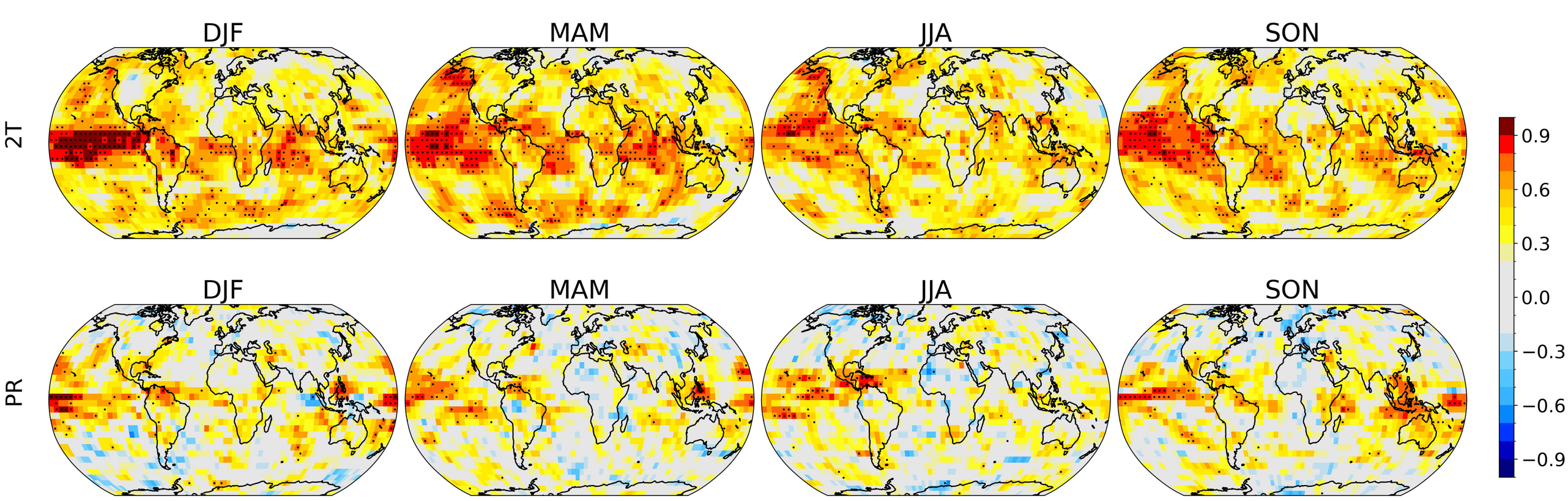
Fig. 6. Regional temperature CRPSS. Temperature forecast skill scores across seasons (DJF, MAM, JJA, and SON) of the three model configurations described in Figure 1, panel D: Global 5°, Regional 5°, and Regional 1°. Metrics are referenced (Ref.) against the climatological forecasts (CLIM) or against the ECMWF's seasonal prediction system (SEAS5). Dots indicate statistical significance that the skill score is positive at the 95% confidence level.

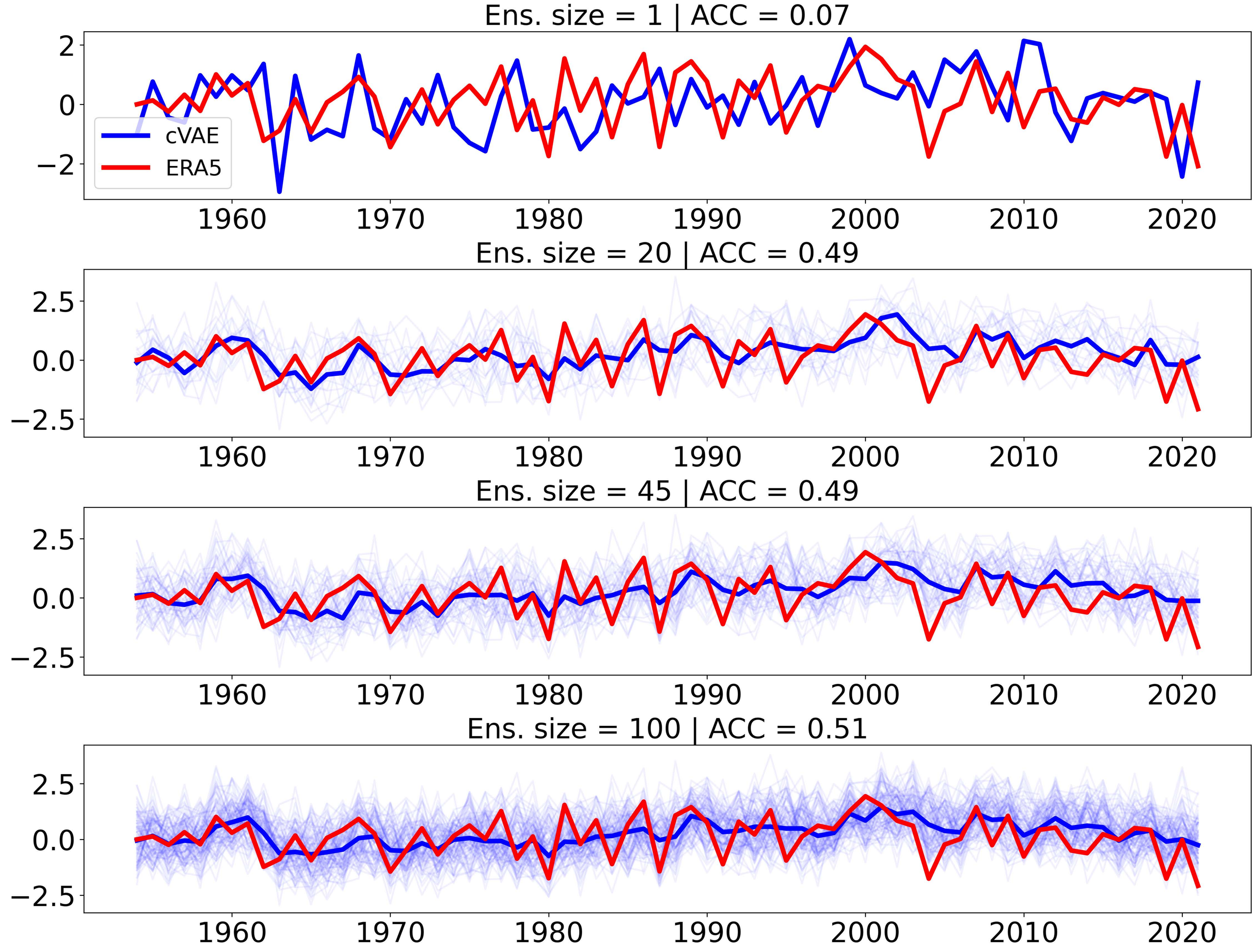
Fig. 7. Regional precipitation CRPSS. Same as Figure 6 but for precipitation predictions.

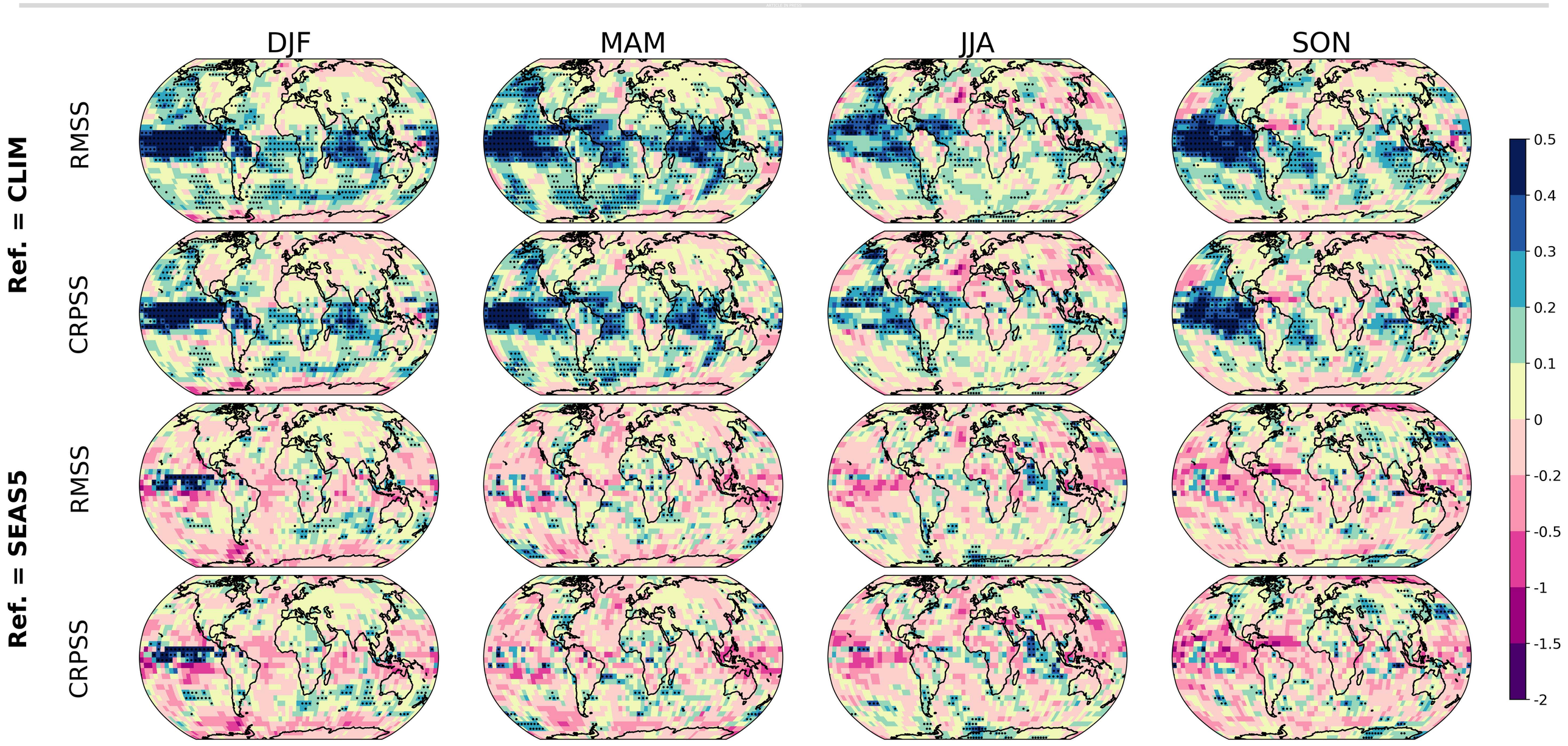
Fig. 8. Niño3.4 teleconnections. Pearson correlation between predicted DJF temperature (2T), precipitation (PR), and 500 hPa geopotential height (ZG500), and the DJF Niño3.4 index (2001–2021). Black dots indicate statistical significance at the 95% confidence level.

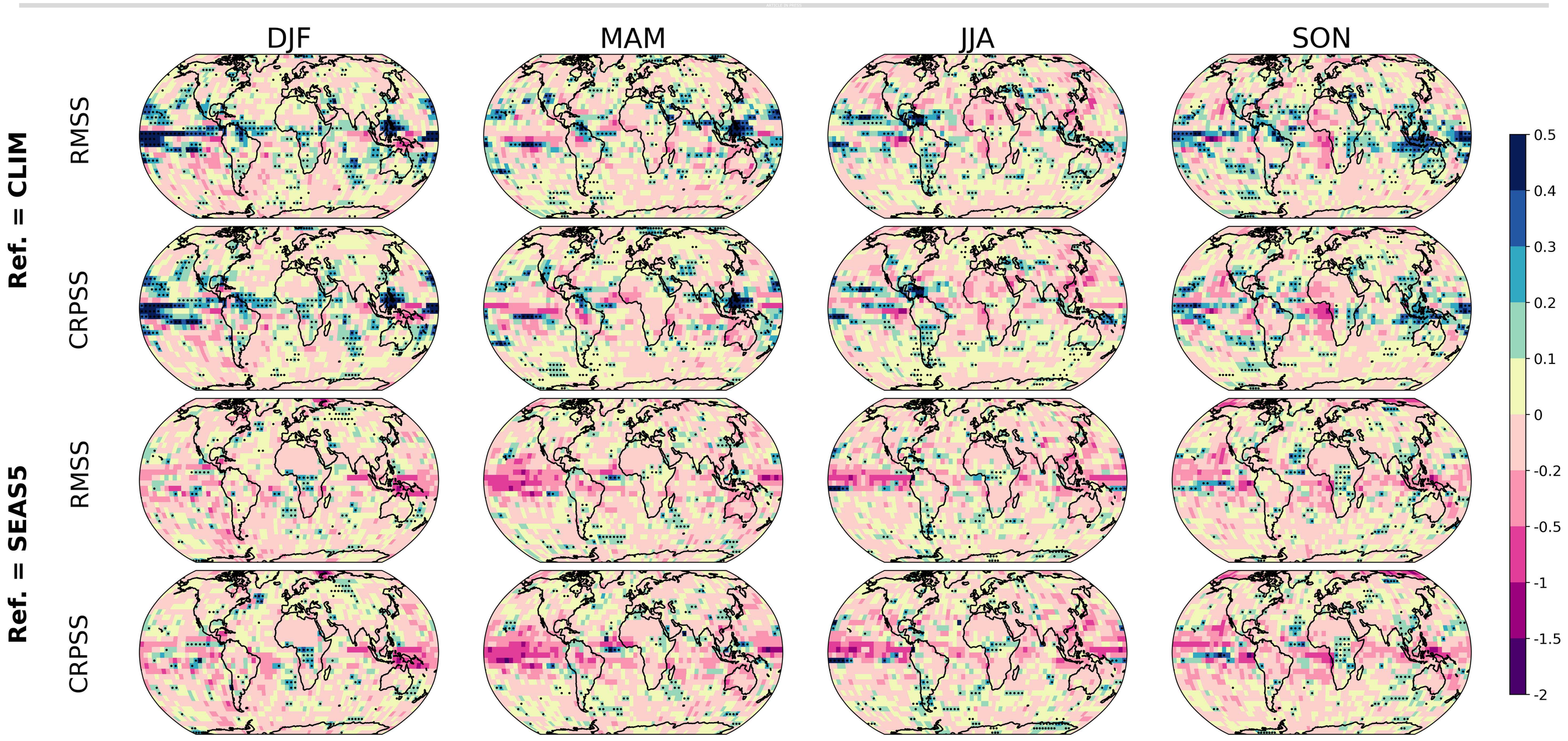
Fig. 9. NAO teleconnections. Pearson correlation between predicted DJF temperature (2T) and precipitation (PR) and the DJF NAO index (2001–2021). Black dots indicate statistical significance at the 95% confidence level.




DJF
MAM
JJA
SON
2T
DJF
MAM
JJA
SON
PR
0.9
0.6
0.3
0.0
−0.3
−0.6
−0.9


Ens. size = 1 | ACC = 0.07
cVAE
ERA5
2
0
−2
1960
1970
1980
1990
2000
2010
2020
Ens. size = 20 | ACC = 0.49
2.5
0.0
−2.5
Ens. size = 45 | ACC = 0.49
Ens. size = 100 | ACC = 0.51


DJF
MAM
JJA
SON
Ref. = CLIM
Ref. = SEAS5
RMSS
CRPSS
RMSS
CRPSS
0.5
0.4
0.3
0.2
0.1
0
-0.2
-0.5
-1
-1.5
-2


DJF
MAM
JJA
SON
Ref. = CLIM
Ref. = SEAS5
RMSS
CRPSS
RMSS
CRPSS
0.5
0.4
0.3
0.2
0.1
0
-0.2
-0.5
-1
-1.5
-2

Ref. = CLIM

Ref. = SEAS5

Global 5° | Regional 5° | Regional 1° | Global 5° | Regional 5° | Regional 1°

DJF

MAM

JJA

SON

0.5
0.4
0.3
0.2
0.1
0
-0.2
-0.5
-1
-1.5
-2

Ref. = CLIM

Ref. = SEAS5

Global 5° | Regional 5° | Regional 1° | Global 5° | Regional 5° | Regional 1°

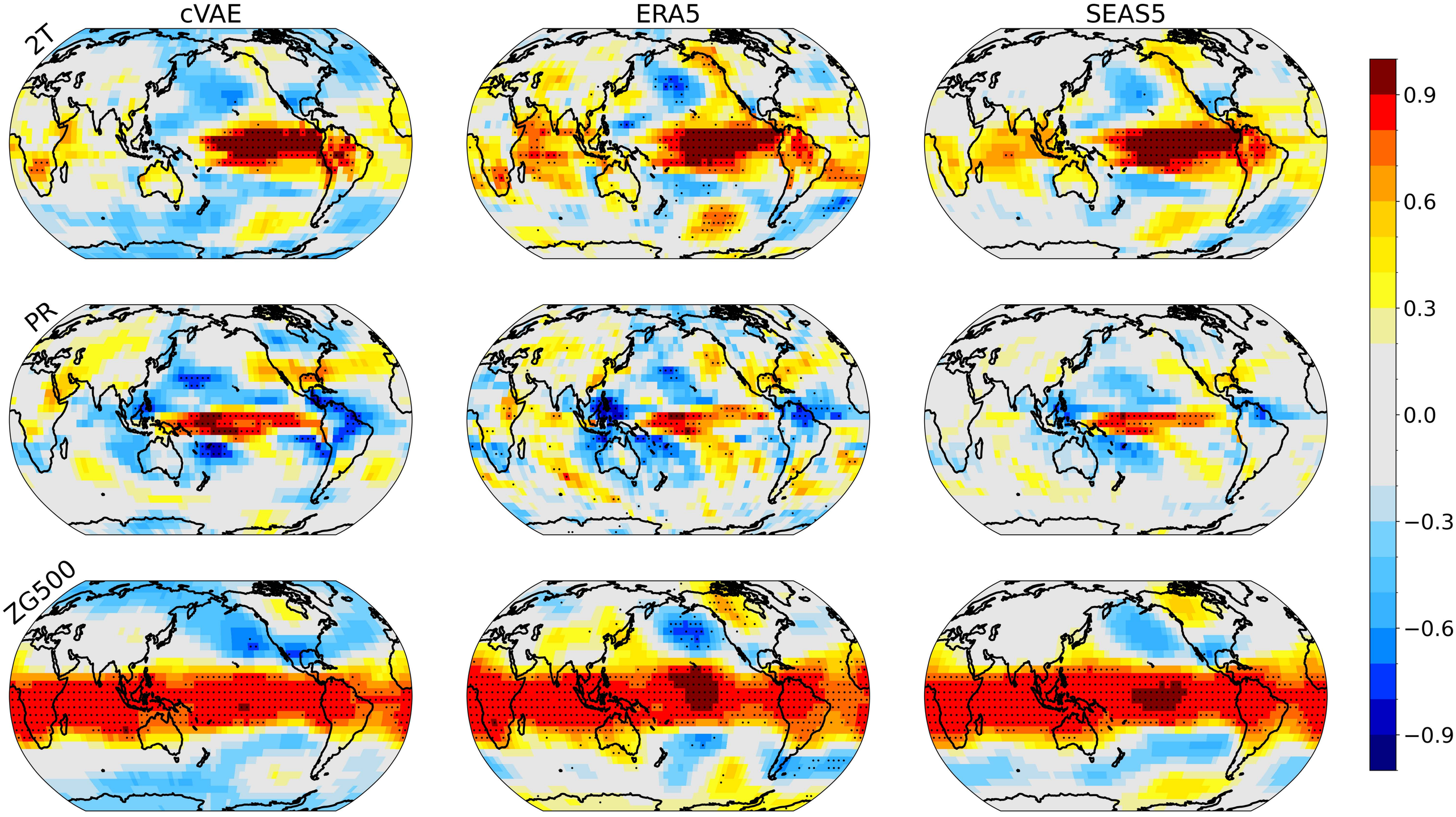
DJF

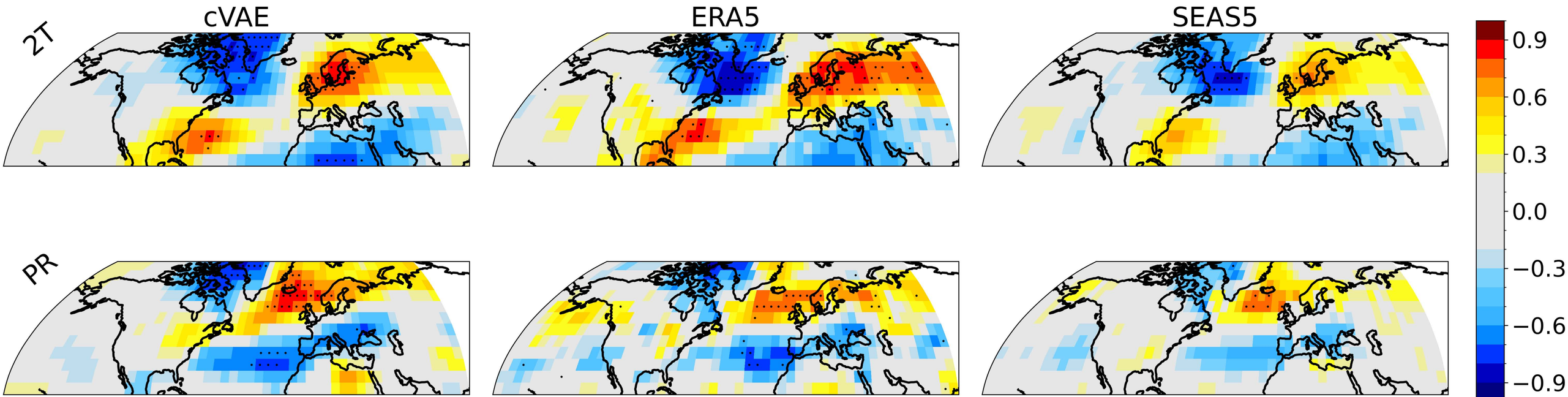
MAM

JJA

SON

0.5
0.4
0.3
0.2
0.1
0
-0.2
-0.5
-1
-1.5
-2


cVAE
ERA5
SEAS5
2T
PR
ZG500
0.9
0.6
0.3
0.0
−0.3
−0.6
−0.9


cVAE
ERA5
SEAS5
2T
PR
0.9
0.6
0.3
0.0
−0.3
−0.6
−0.9