

Statistical downscaling of seasonal forecasts: K-NNs vs CNNs

Author: Lluís Palma García* and Supervisors: Llorenç Lledó¹, lledo@bsc.es;

Carlos Gómez-Gonzalez¹, carlos.gomez@bsc.es; Bernat codina², bcodina@ub.edu

¹*Earth Sciences Department, Barcelona Supercomputing Center (BSC), Barcelona, Spain. and*

²*Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.*

Abstract: Probabilistic information from seasonal forecasts has proven valuable for a wide range of societal sectors such as health, agriculture, or the renewable energy sector, boosting the research community’s interest in making such predictions more skillful and reliable. Existing approaches for seasonal forecasting strongly depend on General Circulation Models (GCMs), which suffer from systematic biases and drifts and present horizontal spatial resolutions that are coarser than those needed for regional or local applications. Hence, arises the need to explore post-processing techniques that can compensate for such limitations. In this study, a Convolutional Neural Network (CNN) architecture tailored for the downscaling of seasonal forecasts at daily resolution is proposed. Using the ECMWF SEAS5 seasonal prediction system, together with the ERA5 and ERA5-Land reanalysis, the proposed methodology is tested and verified against more established methods. The scenario proposed takes November’s initialization targeting the DJF season over the Catalonia region and surroundings. Initial results indicate comparable performance in reducing the model biases and retaining the source skill from the CNN model when compared against benchmark methods—suggesting that the exploration of more complex CNN architectures can yield great potential for seasonal forecasting applications.

I. INTRODUCTION

Seasonal climate forecasts cover the forecasting time scales ranging from a month to slightly longer than a year into the future, filling part of the gap between weather forecasting and climate projections. Probabilistic information from seasonal forecasts has proven valuable for a wide range of societal sectors such as health, agriculture, or the renewable energy sector (*Soret et al. 2019, Torralba et al. 2017, White et al. 2021*), boosting the interest of the research community into making such predictions more skillful and reliable. Existing approaches for seasonal forecasting strongly depend on General Circulation Models (GCMs), which encode physics into dynamical prediction systems coupling atmospheric, ocean and land processes. GCMs suffer from systematic biases and drifts and present horizontal spatial resolutions that are coarser than the ones needed for regional or local applications (*Doblas-Reyes et al. 2013*). Hence, due to the complexity of tackling the sources of such errors, added to the computational cost of increasing GCMs spatial resolution, the need arises to explore post-processing techniques that can compensate for such limitations.

Downscaling techniques are used to mitigate the low spatial resolution of state-of-the-art GCMs via dynamical and statistical modeling. Dynamical downscaling models, also known as Regional Climate Models (RCMs), use nested simulations, with GCMs output as boundary conditions, to model processes that could not be modeled at the spatial resolution of GCMs.

By nature, RCM simulations are computationally expensive and not transferable across different regions. Computational efficiency is a critical aspect for seasonal forecasting because several ensemble members and many past initializations (hindcasts) have to be downscaled to estimate the systematic model errors and the uncertainty associated with initial conditions (*Diez et al. 2005, 2016*). These, added to the fact that RCMs still require post-processing techniques, as they don’t entirely correct the systematic biases present in GCMs (*Manzanas et al. 2018*), are the reasons why dynamical approaches are discarded in most applications at seasonal timescale and also in the present study.

On the contrary, statistical downscaling (SD) (*Marraun and Widmann 2017*) techniques are based on empirical relationships derived between a local observed high-resolution variable of interest (predictand) and one or several lower-resolution variables (predictors). Depending on the origin of the predictors in the training phase, we can differentiate between two main categories ; Model output statistics (MOS), where the predictors are taken directly from the same GCM post-processed, and Perfect prognosis methods (PP), in which both the predictors and the predictand come from observational datasets. Under the PP approach, relationships between coarse and fine observations can be learned at a daily basis and later applied to the daily output of GCMs, adding an extra layer of complexity that is not present in MOS approaches. Moreover, MOS models are limited by GCMs hindcast size (typically around 30 years) and can only be applied to seasonal or monthly aggregated data as temporal correspondence between GCM predictors and local observations is not retained at daily time scales (*Manzanas et al. 2020*).

* Electronic address: lluis.palma@bsc.es

Many studies have explored the pros and cons of both approaches, although most of the literature is oriented towards climate change studies (projections). One example is the VALUE COST action (2012–2015), where a framework for the intercomparison of downscaling techniques for climate projections over Europe was developed (Maraun *et al.* 2015). Regarding seasonal forecasts, the SPECS (Seasonal-to-decadal climate prediction for the improvement of European Climate Services) project analyzed more than 30 publications exploring SD techniques for seasonal forecasting applications (Gutierrez *et al.* 2013). The report concluded that most of the studies worked with seasonal or monthly data instead of daily and that MOS approaches were preferred over PP techniques, with linear regression and analogs algorithms being the most used. Nevertheless, more recent works have explored PP techniques working with daily data (Manzanas *et al.* 2017, 2018, 2020), demonstrating that they can systematically reduce errors in different order moments, ranging from the mean to the extremes.

Although some studies have explored machine learning techniques for SD in the past, including deep-learning (DL) models back in 1998 (Wilby *et al.* 1998), a new resurgence in the field of ML motivated by increasing computational capabilities and improved datasets has occurred over the last decade (Cohen *et al.* 2019, Reichstein *et al.* 2019). Following this line, many studies have explored the adoption of super-resolution (SR, (Wang *et al.* 2019)), a cornerstone task in computer vision aimed at increasing the resolution of images and videos. Most of the proposed methods in SR are based on convolutional neural networks (CNNs). CNNs offer the capability to exploit spatial information by allowing the treatment of high-dimensional data more straightforwardly and bypassing the use of standard feature extraction techniques like Principal Component Analysis (PCA). Recent studies have explored CNNs for climate and weather applications, obtaining improved results over traditional methods (Baño-Medina *et al.* 2020, Höhle *et al.* 2020, Leinonen *et al.* 2020, Vandal *et al.* 2017). Still, very few or no studies propose implementations in the field of seasonal forecasting.

The main aim of this study is to propose a CNN architecture tailored for the downscaling of daily seasonal forecasts under the PP approach and assess its performance compared to more established methodologies. Therefore, the work is divided as follows; Section II describes the seasonal prediction system and reanalysis used. The methodology applied and the verification process are described in Section III. Then, the main results obtained are discussed in Section IV. Finally, the most relevant conclusions are given in Section V.

II. DATA

The present work analyses different SD strategies focused on the Catalonia region and surroundings (2°W , 6°E , 38°N , 46°N), see Figure 1. With this aim, several datasets have been gathered, and in this section, a detailed description of the different products used can be found. Furthermore, Table I shows a summary of the datasets and variables chosen.

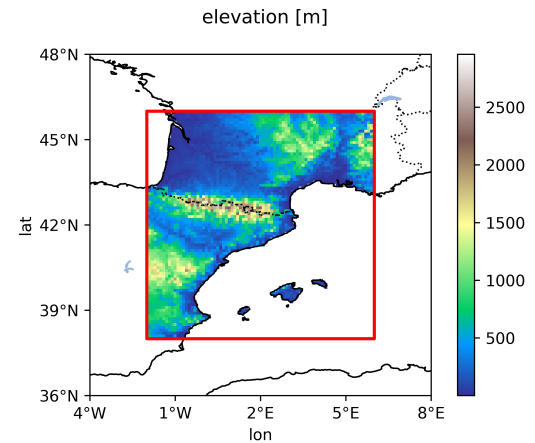


Figure 1: Study domain with topography at 0.1° resolution. The PP models are trained using the whole domain and evaluated within the red box.

A. Seasonal forecasting system

The latest ECMWF seasonal forecasting system SEAS5 (Johnson *et al.* 2019) has been considered for this work. SEAS5 is based on the Integrated Forecast System (IFS) with its atmospheric component coupled to the ocean model NEMO (Nucleus for European Modelling of the Ocean). The operational SEAS5 seasonal forecast comprises an ensemble of 51 members, initialized on the first day of each month and spanning up to 215 days into the future. Besides, a set of hindcasts (retrospective forecasts) are generated for every month of the 1981–2017 period, producing for each initialization a smaller ensemble of 25 members. Although the system is originally run at a horizontal resolution of ~ 36 km (Tco319L91), the final product available at the Copernicus Climate Change Service (C3S) Climate Data Store (CDS) has a regular grid with a spatial resolution of $1^{\circ} \times 1^{\circ}$ (~ 111 km). Thus, the present analysis employs daily forecasts from the November initializations in the 1981–2017 period, and valid for the subsequent DJF seasons (i.e. with a lead time of one month) at 1 degree of resolution.

Dataset	Type	Spatial res.	Period	Temporal res.	Variables	Reference
ERA5-Land	Reanalysis	0.1 ^o (~9 km)	1979-2020	Daily	tas	<i>Muñoz-Sabater et al. (2021)</i>
ERA5	Reanalysis	0.25 ^o (~31 km)	1979-2020	Daily	tas, psl (sfc), ta, hus, g	<i>Hersbach et al. (2020)</i>
SEAS5	GCM	1 ^o (~111 km)	1981-2016	Daily	ua, va (850/500/300 hPa)	<i>Johnson et al. (2019)</i>

Table I: Datasets summary

B. Observational data

Observational data is used both for forecast verification and for the training of the SD model. We differentiate between the low-resolution or coarse source field, the input predictors to train the model, and the high-resolution or fine target field, the final output. On one side, the ECMWF reanalysis ERA5 (*Hersbach et al. 2020*) is used as the low-resolution source dataset. ERA5 presents a spatial grid resolution of 31 km, an hourly time resolution, and 137 vertical levels, conformed using the Numerical Weather Prediction model (IFS Cycle 41r2), improving the data assimilation process to his predecessor, ERA-interim. The current dataset version covers 1979 to 2019, with an extension to 1950 currently in a trial phase.

On the other side, as target resolution, ERA5-Land (*Muñoz-Sabater et al. 2021*) has been employed. ERA5-Land is a re-run of the land component of its counterpart, the ERA5 climate reanalysis, enhancing it with an increase of spatial resolution up to ~9 km and several refinements of the land-surface processes. It runs forced by meteorological fields from ERA5 and without being coupled either to the IFS or the NEMO models, making it computationally inexpensive and, thus, offering the option of quick updates. The 1-hourly temporal frequency from ERA5 is preserved. For both ERA5 and ERA5-Land, daily values covering the whole 1979 to 2019 period are employed in this study. ERA5-Land is used as ground truth in the training process and during the verification, taken as the reference dataset and compared against the downscaled forecasts. Although more precise datasets are available (based on observational measurements), we considered it a good starting point to test the methodologies presented. Furthermore, although being a critical aspect, the observational uncertainty associated with this reanalysis and its implications in the final results have not been tackled in this study.

III. METHODS

A. Standard statistical methods

Besides CNNs, two widely used SD techniques have been implemented to set a proper reference framework. The first one falls within the category of MOS techniques and consists of a combination of a nearest-neighbor in-

terpolation followed by a Simple Bias Correction (SBC). Under the assumption that both the reference and the predicted distributions are gaussian, the SBC method creates predictions with the same mean and standard deviation as the reference field (*Torrallba et al. 2017*). The scheme is summarised by the following expression:

$$y_{ij} = (x_{ij} - \bar{x}) \frac{\sigma_{ref}}{\sigma_e} + \bar{o} \quad (1)$$

Where x_{ij} is the seasonal average for each member j and each year i of the hindcast, and y_{ij} is the resulting bias-adjusted seasonal average. Thus, \bar{x} and \bar{o} denote the climatologies of the hindcast ensemble mean and the reference datasets, and $\frac{\sigma_{ref}}{\sigma_e}$ the ratio between the standard deviation of the reference dataset σ_{ref} and the interannual standard deviation of the ensemble members σ_e . As mentioned, this technique is applied directly over seasonal averages of the predictand field without considering large-scale predictors, making it a very straightforward and computationally cheap methodology that can be used over different regions and domains.

Alternatively, a PP method has been implemented based on the popular analog technique (*Lorenz 1969*), which infers the downscaled forecast from a set of analog situations selected from daily historical observational data. The search of the analogous fields is performed using the Euclidean distance, which estimates the similarity between large-scale forecast fields and their counterparts from historical reanalysis data. Thus, once the analogous days are selected, the resulting predictand ERA5-Land fields are averaged and taken as the new forecast. This approach is also known outside the climate community as an iteration of the k-nearest neighbours (K-NNs) algorithm (*James et al. 2021*), considering each grid point as a feature and getting a multioutput. This algorithm is applied over each forecast day and every ensemble member of SEAS5, and to accommodate the spatial resolution to ERA5's resolution, nearest-neighbor interpolation has been performed to the latter. On the same line, the climate community has widely tested similar analog approaches that offer decent results for seasonal forecasting applications (*Manzanas et al. 2017, 2018, Nikulin et al. 2018*).

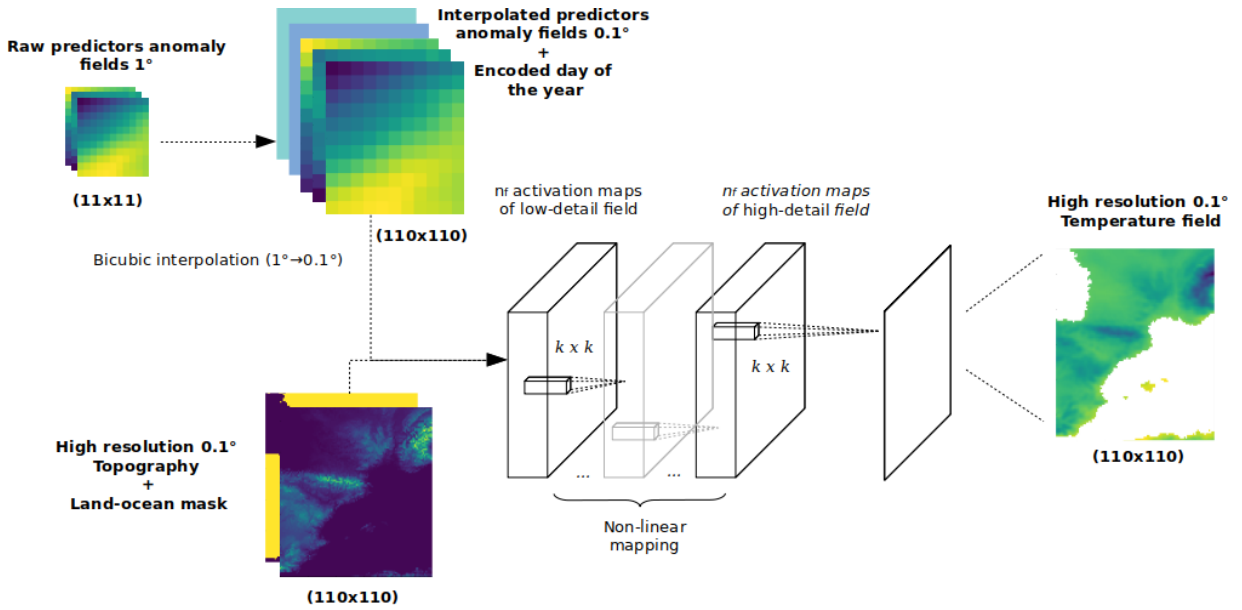


Figure 2: Augmented SRCNN Architecture. Low resolution daily predictors fields are interpolated to the grid of the HR field via bicubic interpolation and concatenated with HR topography and land-ocean masks. After passing through the different convolutional steps, the daily temperature field is obtained with HR features.

B. Convolutional Neural Networks (CNNs)

As mentioned earlier, under the name of SR algorithms, many strategies have been proposed in computer vision to learn the mapping between low and high-resolution images. On this basis, some attempts have already been made to downscale daily Essential Climate Variables (ECVs) from projections (Baño-Medina et al. 2021, Vandal et al. 2018), mainly oriented on applying Super-Resolution Convolutional Neural Networks (SRCNNs) (Dong et al. 2016) via the PP approach. This work presents a similar method, based on a similar architecture, and aimed at downscale daily mean temperature from seasonal forecasts.

Traditional Fully Connected Neural Networks (FCNNs) are based on the interconnection of numerous neurons through multiple stacked layers. A set of weights W and biases B learned from data through an optimization process and applied to a non-linear mathematical transformation characterize each connection. The combination of layers and neurons offers the capability to approximate complex non-linear functions. In a hypothetical implementation of an FCNN-based algorithm for downscaling temperature fields, the different raw predictor’s fields would have been transformed into a 1-dimensional vector and then connected to each neuron of the first layer. For large spatial domains with high spatial resolution, this would imply a drastic increase of the parameters to be optimized, growing as a function of the number of grid cells and neurons, and carrying the risk of heavily overfitting the data.

Under this scenario, CNNs have long surpassed traditional FCNNs in computer vision and, more specifically, in the field of SR (Dong et al. 2016, Wang et al. 2019). They exploit convolutions to extract information from data by applying a set of 3D (latitude, longitude, channel) sliding kernels or filters across the input fields. Thus, as kernels are shared for each convolutional layer over the whole spatial domain, CNNs imply a substantial reduction in the degrees of freedom, making a model less prone to overfit and with enhanced computational efficiency.

The proposed implementation applies multiple convolutional layers to one or multiple pre-upsampled (i.e., coarse to fine interpolation) LR fields via bicubic interpolation; this approach is known as pre-upsampling super-resolution. Hence, as a first step, a convolution is applied to the interpolated fields X by convolving the weights and biases of each kernel and passing the result through a Rectified Linear Unit (Relu) function that serves as non-linear transformation (see Equation 3 where “ $*$ ” is the convolution operation). This process is applied, taking the input of the previous layer $F_{l-1}(X)$ as many times as layers (L) present in the model, and finishes with a final convolutional layer obtaining as output the deterministic daily temperature field $F(X)_L$, see Equation 4.

$$F_0(\mathbf{X}) = \mathbf{X} \quad (2)$$

$$F_l(\mathbf{X}) = \max(0, W * F_{l-1}(\mathbf{X}) + B) \quad (3)$$

$$F_L(\mathbf{X}) = W * F_{L-1}(\mathbf{X}) + B \quad (4)$$

On the first layer ($l = 1$), W_1 consists of n_{f_1} filters with shape $k \times k \times c$, where k is the filter size and c is the number of predictors passed as input whereas in the subsequent layers the kernels present the shape $k \times k \times n_{f_{l-1}}$. Thus, to maximize the network’s performance, a screening procedure has to be executed to find the most optimal combination of layers and filters. Several permutations of the number of layers (10, 20, 30, 40 & 60) and the number of filters per layer (16, 32, 64 & 128) were tested (not shown), obtaining the best result for the architecture presenting 20 convolutional layers with 64 filters each. Although other architectures presented in the literature vary the number of filters over each layer, to seek simplicity, this variable has been fixed over all the layers as no specific criteria for selecting those variations have been found. Similarly, a filter size ($k \times k$) of 3×3 is picked, in line with the work done by (Baño-Medina et al. 2020). Therefore, the resulting end-to-end mapping requires the optimization of the parameters $\Theta = \{W_1, \dots, W_L, B_1, \dots, B_L\}$. With this purpose, the Adam optimizer (Kingma and Ba 2014) and the Mean Absolute Error (MAE) as loss function have been selected, where the regressor fields from the SRCNN are taken as inputs X_i , and the HR predictand fields are taken as labels Y_i - following the expression:

$$\mathcal{L}(\Theta) = \frac{1}{n} \sum_{i=1}^n |F(X_i; \Theta) - Y_i| \quad (5)$$

Figure 2 summarises the architecture of the SRCNN. During training, ERA5 predictors fields are first coarsened to SEAS5 1° resolution; the same fields are interpolated using bicubic interpolation and concatenated with High-resolution topography and land-ocean masks over the "channel" dimension. The predictor’s fields are passed as daily anomalies of each grid point independently to avoid systemic biases from the mismatch between the dataset used for training (ERA5) and inference (SEAS5). A similar approach has been tested in (Baño-Medina et al. 2021). In addition, to avoid losing seasonal cycle information, the day of the year is encoded using the cosine and sine transformations and concatenated to the input tensor. Then, the fields are convolved over the many layers described earlier, and the resulting output is the high-res temperature field. Padding is applied to preserve the size of the (latitude, longitude) tensors after each convolutional layer. Therefore, training takes ERA5 and ERA5-Land daily fields from 1981 to 2016 (13,148 samples), reserving 2017 to 2020 as a holdout set (1460 samples), with a batch size of 32 samples, and through 300 epochs with a learning rate ranging between $1e-4$ and $1e-5$. The model has been implemented using the Tensorflow/Keras framework and trained in the CTE-POWER cluster of the Barcelona Supercomputing Center, which mounts Nvidia’s V100 GPUs.

C. Predictors selection

The selection of predictors for the SD of seasonal forecasts under the PP approach is not trivial, and two main factors need to be taken into account. On the first hand, the capabilities of the chosen predictors representing at a daily time scale the local field of interest (temperature in our case) have to be assessed. Thus, Figure 8 shows the daily correlation between the selection of ERA5 predictors and ERA5-Land temperature. Fields are ranked by the absolute mean over the predictor domain, as strong positive and negative correlations suggest a solid physical link. Unsurprisingly, fields like the temperature at surface (tas) and 850 hPa (ta850), or the specific humidity at the same level (hus850), are among the fields better correlated with daily temperature at a local scale.

Furthermore, (Manzanas et al. 2018) pointed out that, under the context of seasonal forecasting, it is essential to consider the skill of our prediction system in representing such large-scale predictors. Hence, Figure 9 shows the Anomaly Correlation Coefficient (ACC) of DJF seasonal means between the fields of the SEAS5 forecast initialized in November and the observed fields from ERA5. Again, the mean absolute value has been computed over the predictor’s domain (indicated with a red box). Strikes the lack of predictability in the SW European region for many of the variables considered, which is a well-known problem in the community of Seasonal forecasting. Not in line with the results obtained in Figure 8, the predictors with higher skill are primarily within high altitude (300 hPa) fields and those of most interest (ta850, hus850, or tas) present poor skill. Thus, a trade-off has to be made between both metrics discussed here, and as a first approach, we compute an index based on the multiplication of both indices and rank each variable based on it. Results can be shown in Figure 3, a vertical red line is indicating all the predictors discarded as a results of the analysis (hus500 onwards).

	ta300	tas	g300	hus300	ta500	hu500	g500	ua500	hus500	ua500	g500	ta850	psi	va300	va500	va850	ua850
Skill * Daily Corr.	0.17	0.13	0.13	0.11	0.11	0.08	0.07	0.06	0.05	0.04	0.04	0.04	0.03	0.02	0.02	0.01	0.01
Daily Corr.	0.32	0.96	0.52	0.44	0.51	0.35	0.34	0.24	0.62	0.29	0.50	0.61	0.16	0.28	0.22	0.18	0.16
Skill	0.53	0.14	0.24	0.26	0.21	0.22	0.22	0.25	0.07	0.15	0.08	0.06	0.20	0.07	0.09	0.07	0.07

Figure 3: According to predictive skill and daily correlation, top ranking predictors are ordered from left to right. The vertical red line marks the cut between the selected predictors and the discarded ones.

D. Forecast verification

To assess the results of the different downscaling methods, we analyze a set of metrics over the hindcast period (1981–2016). As verifying the forecasts with the same period used for training can lead to the appearance of spurious predictive skill, and to evaluate the potential use of such models in an operational forecasting context, each initialization of both the SBC and K-NNs methods has been produced under a Leave-One-year-Out (LOO) cross-validation scheme. Thus, the bias and the Anomaly Correlation Coefficient (ACC) are computed over the ensemble mean to explore the deterministic output of the methods. Similarly, the fair Ranked Probability Skill Score (fRPSS) is computed to verify the probabilistic skill of the forecasts providing probabilities for the above normal, normal, and below normal categories. Furthermore, to assess the performance of predicting extreme events, the biases of the 10th and 90th percentiles and the fair Brier Skill Score (fBSS) for the same categories are also assessed. All the metrics above are computed over the DJF seasonal means, and fair versions of the probabilistic skill metrics are given to account for the differences in the hindcast and forecast ensemble sizes, adapting the results to a theoretical real-time operational forecasting scenario.

IV. RESULTS

In this section, the methods presented are analyzed and discussed. The chapter is divided into two main blocks. In the first one (Section IV A), the methodologies implemented are tested using only surface temperature (tas) as a predictor. On the contrary, in the second section (Section IV B), the same methods are tested but assessing the impact of adding multiple predictors, following the criteria described in Section III C.

A. Temperature as only predictor

Figure 4 shows the verification results obtained using only surface temperature as predictor. The raw SEAS5 forecast (RAW) interpolated to the ERA5-Land grid using nearest-neighbor interpolation has also been added as a benchmark. In general, all the methods drastically reduced the biases present in the SEAS5 raw model, and the SBC clearly outperforms the other methods. This is not a surprise, as the SBC method is specifically focused on reducing these biases. On the same line, comparing the Analogs method and the CNN, the latter provides better performance on this task and gets closer to the SBC method, especially reducing the 10th and 90th percentiles (P10 and P90) biases. If we look at probabilistic skill metrics (fRPSS, fBSS10, and fBSS90), it strikes the very low skill values present in the raw forecast even after applying all the SD methods and how very few points

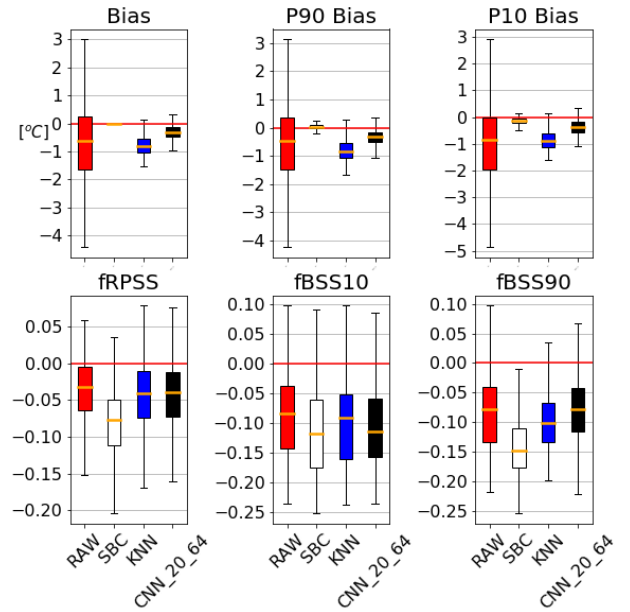


Figure 4: Verification results obtained using only surface temperature as predictor. Each panel showcases a different metric, and each box-plot represents the spatial distribution of each of the methods considered in Section III (the boxes correspond to the 25–75th percentile interquartile range, whereas the whiskers cover the min-max range). The orange line corresponds to the mean value of the sample set.

manage to perform better than the climatology. Under this scenario, we notice a deterioration of the skill of the forecast when applying the SBC process, which counterpoints the bias-reduction capabilities of the methods. Besides, the CNN and Analogs techniques are able to maintain the mince skill levels present in the raw prediction without significant differences between the two.

The same results are also shown in map form in Figure 5. The generalized reduction performed by the SBC over the SEAS5 biases is noticeable, in addition to the low skill present in the region, where very few regions perform above the climatology. Taking special attention to the CNN’s result, a noisy pattern can be distinguished over the first three plots, similar patterns have been experienced in previous implementations of CNNs for SD, but no specific clue has been found on its source, and it is considered a critical point to be addressed in the future. Thus, these results are well in-line with other works tackling the performance of PP techniques (*Manzanas et al.* 2018, 2020, *Nikulin et al.* 2018); as working under the PP framework using the same field as predictor does not allow the improvement of the prediction skill, and only a reduction of the biases while maintaining (as much as possible) the source predictable skill can be expected.

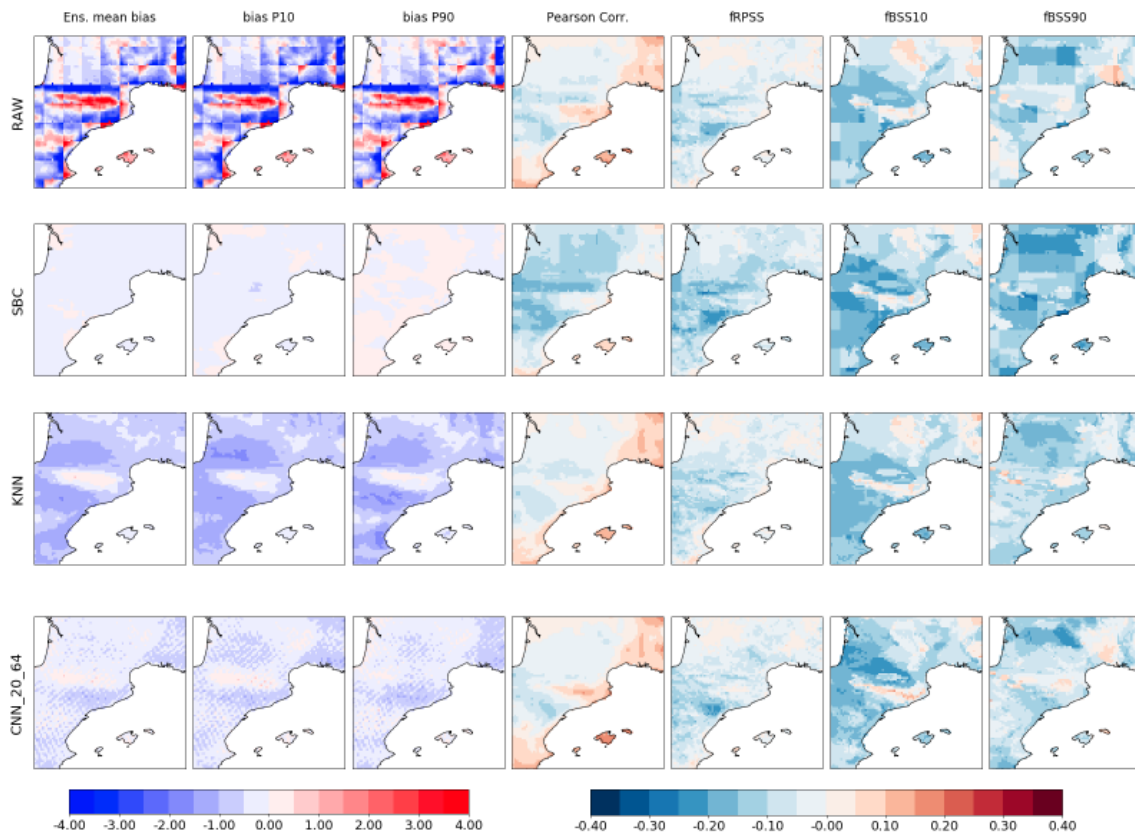


Figure 5: Maps of the spatial results obtained for the different metrics described in Section III D and the multiple methods presented. The raw SEAS5 forecast (RAW) interpolated to the ERA5-Land grid using nearest-neighbour interpolation is added as benchmark.

B. Adding extra large-scale predictors

As mentioned in the previous section, PP methods taking as only predictor the same predictand variable (tas in this case), cannot increase the skill of seasonal predictions. This is mainly due to their training being performed using observational datasets, assuming that the input predictors are perfect and, hence, not having any additional source of predictive skill. Thus, this approach cannot perform any correction to increase the forecast's skill by itself and, thus, depends entirely on the skill of the predictors used. With this in mind, (Manzanas *et al.* 2018, 2020) pointed out that, under the context of seasonal forecasting, the skill for a target variable might be increased if the skill of the variables used as predictors results to be higher than the one of the target field. Thus, starting with Section III C analysis, several permutations of different variables were tested.

Figure 6 shows the results obtained for three configurations combining temperature fields at various levels (surface, 300 hPa, and 500 hPa), followed by the geopotential height at 300 hPa and the specific humidity at the same level. Those predictors were the best for our region and season, offering the best trade-off between skill and representativeness of local temperature (see figure 3). Simi-

larly to figure 4, each panel showcases a different metric, and each box-plot is one of the methods considered in section III. Blue boxes correspond to the different K-NN implementations, while black boxes indicate the CNN's ones. The dashed grey line divides the implementation using only tas as a predictor (discussed in the previous section) from the new ones. In this second scenario, the fBSS10 metric has been exchanged by the ensemble mean correlation (ACC), as the first one was not showing significant differences between the methods compared.

In general terms, K-NN implementations using multiple variables as predictors tend to produce lower biases than CNNs, which do not seem to improve. The K-NN implementation employing all the predictors considered surpasses the best performer CNN implementation in terms of biases. However, while none of the methods can outperform the SBC in those terms, all the methods outperform the SBC when considering skill metrics (CORR, fBSS90, and fRPSS). Similar to the results obtained in the bias metrics, CNNs do not seem to improve when adding extra predictors, and K-NNs offer better results, especially when considering all the predictors. The same results are also shown in map form in Figure 7. An apparent enhancement of the K-NN methods can be seen when adding multiple predictors,

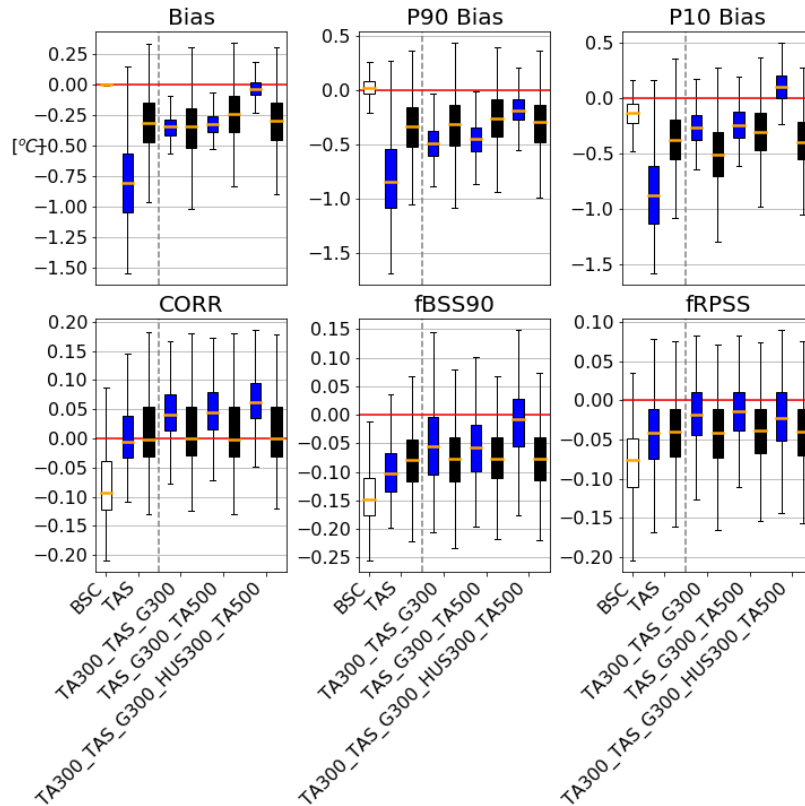


Figure 6: Verification results were obtained using only multiple predictors. Each panel showcases a different metric, and each box-plot is one of the methods considered in Section III (the boxes correspond to the 25–75th percentile interquartile range, whereas the whiskers cover the min-max range). The orange line corresponds to the mean value of the sample set. Blue boxes correspond to the different Analogs implementations, while black boxes indicate the CNN’s ones. SBC of the raw tas field is added as a benchmark (white box).

both in terms of biases and skill, while the CNNs cannot improve with the extra predictors. Still, the enhancement is modest due to the limited skill of all the included predictors. However, both the CNN and K-NN implementations manage to reduce the biases present in the RAW forecast while not reducing the skill, which is a clear advantage over the SBC method.

V. CONCLUSIONS

A new resurgence in the field of machine learning motivated by increasing computational capabilities and improved datasets has occurred over the last decade (Cohen *et al.* 2019, Reichstein *et al.* 2019). Following this trend, many studies have explored the adoption of Convolutional Neural Networks (CNNs) for climate and weather applications, obtaining improved results over traditional methods (Baño-Medina *et al.* 2020, Höhle *et al.* 2020, Leinonen *et al.* 2020, Vandal *et al.* 2017). Still, very few or no studies propose implementations in the field of seasonal forecasting.

In this study, we propose a CNN architecture trained under the Perfect Prognosis (PP) approach and suited for the downscaling of seasonal temperature fields, reaching a 10x increase in spatial resolution. The proposed architecture is evaluated against two well-known statistical downscaling techniques: the Analogs algorithm and a Simple Bias Correction (SBC). Thus, several metrics comparing each model’s deterministic and probabilistic output against observations have been computed to access the quality of the different implementations.

When using the temperature at the surface both as predictand and as the only predictor, results show how all the methods drastically reduced the biases present in the SEAS5 coarse model, where the SBC outperforms the other methods, followed by the CNN implementation and the k-nearest neighbors (K-NNs) algorithm. However, in terms of forecast skill and association, none of the methods can increase the low performance of the raw forecasts. While the SBC deteriorates the skill, counterpointing the strong bias-reduction capabilities of the method, the CNN and K-NN algorithms manage to retain the original skill.

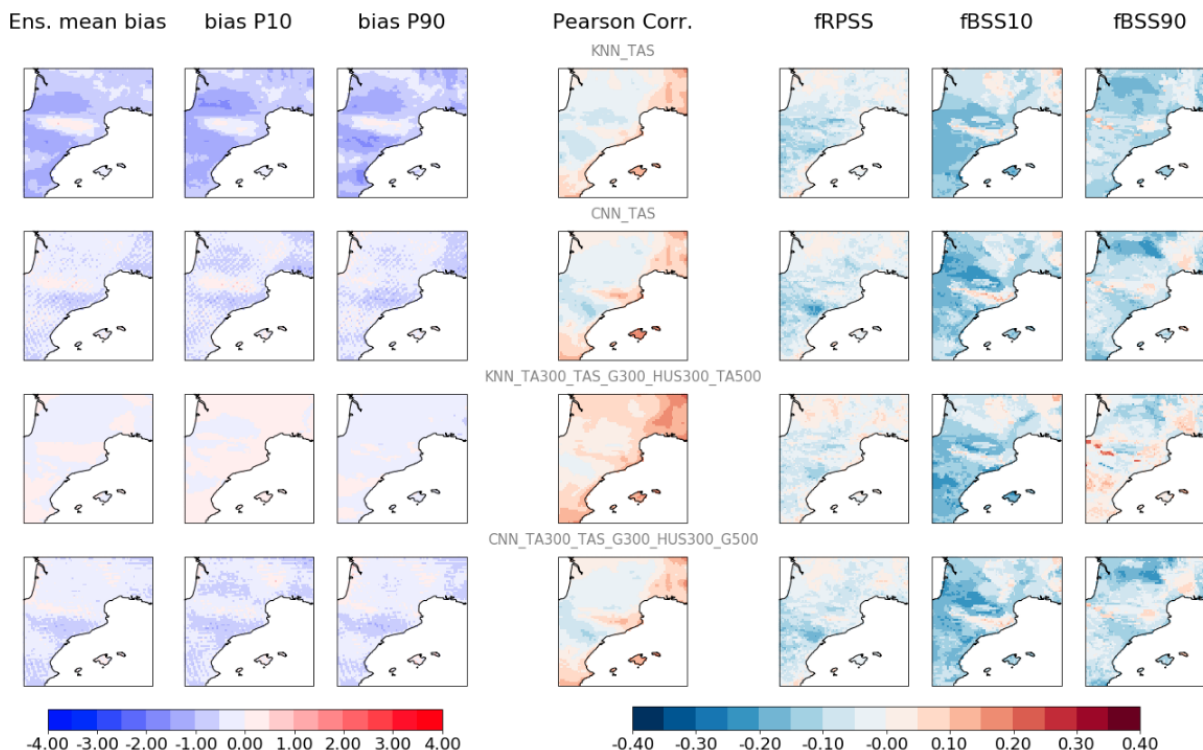


Figure 7: Maps of the spatial results obtained for the different metrics described in Section III D and the multiple methods presented combining multiple predictors. Only the best performer combination of predictors is shown (see Figure 6).

Furthermore, when considering multiple predictors at different pressure levels, the limited skill found in SEAS5 predicting those fields for the studied region represents a significant handicap in the capabilities of the different methods. Despite the limitations, adding multiple predictors improves the reduction of the mean biases in the case of the K-NN methods, performing better than the CNNs, which do not seem to improve under this new configuration. Still, the CNNs significantly reduce the mean biases while retaining the skill found in the raw model and getting very close in performance to the best K-NN model. It should be noted that the CNN architecture employed here is relatively simple, it has not been trained with a dedicated loss function that would optimize over the metrics used in this study, and it is far from the current state-of-the-art techniques in computer vision. Therefore, exploring more complex architectures, training procedures, and tailored loss functions still encloses great potential for the field of SD. Nevertheless, from a practical point of view, the presented CNN architecture can be extrapolated to different regions and domains, where higher seasonal skill is found, offering great potential due to their flexibility and computational efficiency. In addition, due to the poor skill present in current seasonal prediction systems in the region of study, Model output statistics (MOS) approaches merged with deep learning algorithms are also potential lines of investigation and will be explored in the future.

ACKNOWLEDGMENTS

I would like to thank my supervisors Llorenç Lledó, Carlos Gómez González, and Bernat Codina for their valuable support and discussions. I also acknowledge Pierre-Antoine Bretonnière and Margarida Samsó for their invaluable help gathering and processing the datasets used. Last but not least, I express my sincere gratitude to my friends and family, whose unconditional support was vital in these strange times.

REFERENCES

Baño-Medina, J., R. Manzanar, and J. M. Gutierrez, Configuration and intercomparison of deep learning neural models for statistical downscaling, *Geoscientific Model Development*, 13(4), 2109–2124, doi:10.5194/gmd-13-2109-2020, 2020.

Baño-Medina, J., R. Manzanar, and J. M. Gutiérrez, On the suitability of deep convolutional neural networks for continental-wide downscaling of climate change projections, *Climate Dynamics*, 57(11-12), 2941–2951, doi:10.1007/s00382-021-05847-0, 2021.

Cohen, J., D. Coumou, J. Hwang, L. Mackey, P. Orenstein, S. Totz, and E. Tziperman, S2S reboot: An argument for greater inclusion of machine learning in subseasonal to seasonal forecasts, *Wiley Interdisciplinary Reviews: Climate Change*, 10(2), 1–15, doi:10.1002/wcc.567, 2019.

- Díez, E., C. Primo, J. A. García-Moya, J. M. Gutiérrez, and B. Orfila, Statistical and dynamical downscaling of precipitation over Spain from DEMETER seasonal forecasts, *Tellus Series A: Dynamic Meteorology and Oceanography*, 2005, 57(3), 409–423, 57(3), 409–423, doi:10.1111/J.1600-0870.2005.00130.X, 2005.
- Díez, E., B. Orfila, M. D. Frías, J. Fernández, A. S. Cofiño, and J. M. Gutiérrez, Downscaling ECMWF seasonal precipitation forecasts in Europe using the RCA model, <http://dx.doi.org/10.1111/j.1600-0870.2011.00523.x>, 63(4), 757–762, doi:10.1111/J.1600-0870.2011.00523.X, 2016.
- Doblas-Reyes, F. J., J. García-Serrano, F. Lienert, A. P. Biscas, and L. R. Rodrigues, Seasonal climate predictability and forecasting: status and prospects, *Wiley Interdisciplinary Reviews: Climate Change*, 4(4), 245–268, doi:10.1002/WCC.217, 2013.
- Dong, C., C. C. Loy, K. He, and X. Tang, Image Super-Resolution Using Deep Convolutional Networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2), 295–307, doi:10.1109/TPAMI.2015.2439281, 2016.
- Gutierrez, J. M., J. Bedia, R. Benestad, and C. Pagé, Local predictions based on statistical and dynamical downscaling, http://www.specs-fp7.eu/sites/default/files/ui/SPECS_D52.1.pdf, SPECS project deliverable 5.2.1, 2013.
- Hersbach, H., et al., The ERA5 global reanalysis, *Quarterly Journal of the Royal Meteorological Society*, 146(730), 1999–2049, doi:10.1002/qj.3803, 2020.
- Höhlein, K., M. Kern, T. Hewson, and R. Westermann, A comparative study of convolutional neural network models for wind field downscaling, *Meteorological Applications*, 27(6), doi:10.1002/MET.1961, 2020.
- James, G., D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, Springer US, doi:10.1007/978-1-0716-1418-1, 2021.
- Johnson, S. J., et al., SEAS5: The new ECMWF seasonal forecast system, *Geoscientific Model Development*, 12(3), 1087–1117, doi:10.5194/gmd-12-1087-2019, 2019.
- Kingma, D. P., and J. L. Ba, Adam: A Method for Stochastic Optimization, *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2014.
- Leinonen, J., D. Nerini, and A. Berne, Stochastic Super-Resolution for Downscaling Time-Evolving Atmospheric Fields with a Generative Adversarial Network, *IEEE Transactions on Geoscience and Remote Sensing*, 59(9), 7211–7223, doi:10.1109/TGRS.2020.3032790, 2020.
- Lorenz, E. N., ATMOSPHERIC PREDICTABILITY AS REVEALED BY NATURALLY OCCURRING ANALOGUES, *Journal of the Atmospheric Sciences*, doi:10.1175/1520-0469(1969)26<636:aparbn>2.0.co;2, 1969.
- Manzanas, R., A. Lucero, A. Weisheimer, and J. M. Gutiérrez, Can bias correction and statistical downscaling methods improve the skill of seasonal precipitation forecasts?, *Climate Dynamics* 2017 50:3, 50(3), 1161–1176, doi:10.1007/S00382-017-3668-Z, 2017.
- Manzanas, R., J. M. Gutiérrez, J. Fernández, E. van Meijgaard, S. Calmanti, M. E. Magariño, A. S. Cofiño, and S. Herrera, Dynamical and statistical downscaling of seasonal temperature forecasts in Europe: Added value for user applications, *Climate Services*, 9, 44–56, doi:10.1016/j.cliser.2017.06.004, 2018.
- Manzanas, R., J. M. Gutiérrez, J. Bhend, S. Hemri, F. J. Doblas-Reyes, E. Penabad, and A. Brookshaw, Statistical adjustment, calibration and downscaling of seasonal forecasts: a case-study for Southeast Asia, *Climate Dynamics* 2020 54:5, 54(5), 2869–2882, doi:10.1007/S00382-020-05145-1, 2020.
- Maraun, D., and M. Widmann, Statistical Downscaling and Bias Correction for Climate Research, *Statistical Downscaling and Bias Correction for Climate Research*, doi:10.1017/9781107588783, 2017.
- Maraun, D., M. Widmann, J. M. Gutiérrez, S. Kotlarski, R. E. Chandler, E. Hertig, J. Wibig, R. Huth, and R. A. Wilcke, VALUE: A framework to validate downscaling approaches for climate change studies, *Earth’s Future*, 3(1), 1–14, doi:10.1002/2014EF000259, 2015.
- Muñoz-Sabater, J., et al., ERA5-Land: A state-of-the-art global reanalysis dataset for land applications, *Earth System Science Data*, 13(9), 4349–4383, doi:10.5194/ESSD-13-4349-2021, 2021.
- Nikulin, G., et al., Dynamical and statistical downscaling of a global seasonal hindcast in eastern Africa, *Climate Services*, 9, 72–85, doi:10.1016/j.cliser.2017.11.003, 2018.
- Reichstein, M., G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat, Deep learning and process understanding for data-driven Earth system science, *Nature* 2019 566:7743, 566(7743), 195–204, doi:10.1038/s41586-019-0912-1, 2019.
- Soret, A., V. Torralba, N. Cortesi, I. Christel, L. Palma, A. Manrique-Suñén, L. Lledó, N. González-Reviriego, and F. J. Doblas-Reyes, Sub-seasonal to seasonal climate predictions for wind energy forecasting, *Journal of Physics: Conference Series*, 1222(1), 012,009, doi:10.1088/1742-6596/1222/1/012009, 2019.
- Torralba, V., F. J. Doblas-Reyes, D. MacLeod, I. Christel, and M. Davis, Seasonal Climate Prediction: A New Source of Information for the Management of Wind Energy Resources, *Journal of Applied Meteorology and Climatology*, 56(5), 1231–1247, doi:10.1175/JAMC-D-16-0204.1, 2017.
- Vandal, T., E. Kodra, and A. R. Ganguly, Intercomparison of Machine Learning Methods for Statistical Downscaling: The Case of Daily and Extreme Precipitation, *Theoretical and Applied Climatology*, 137(1-2), 557–570, doi:10.1007/s00704-018-2613-3, 2017.
- Vandal, T., E. Kodra, and A. R. Ganguly, Intercomparison of machine learning methods for statistical downscaling: the case of daily and extreme precipitation, *Theoretical and Applied Climatology* 2018 137:1, 137(1), 557–570, doi:10.1007/S00704-018-2613-3, 2018.
- Wang, Z., J. Chen, and S. C. Hoi, Deep Learning for Image Super-resolution: A Survey, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10), 3365–3387, doi:10.1109/TPAMI.2020.2982166, 2019.
- White, C. J., et al., Advances in the application and utility of subseasonal-to-seasonal predictions, *Bulletin of the American Meteorological Society*, -1(aop), 1–57, doi:10.1175/BAMS-D-20-0224.1, 2021.
- Wilby, R. L., T. M. Wigley, D. Conway, P. D. Jones, B. C. Hewitson, J. Main, and D. S. Wilks, Statistical downscaling of general circulation model output: A comparison of methods, *Water Resources Research*, 34(11), 2995–3008, doi:10.1029/98WR02577, 1998.

Appendix A: Additional Figures

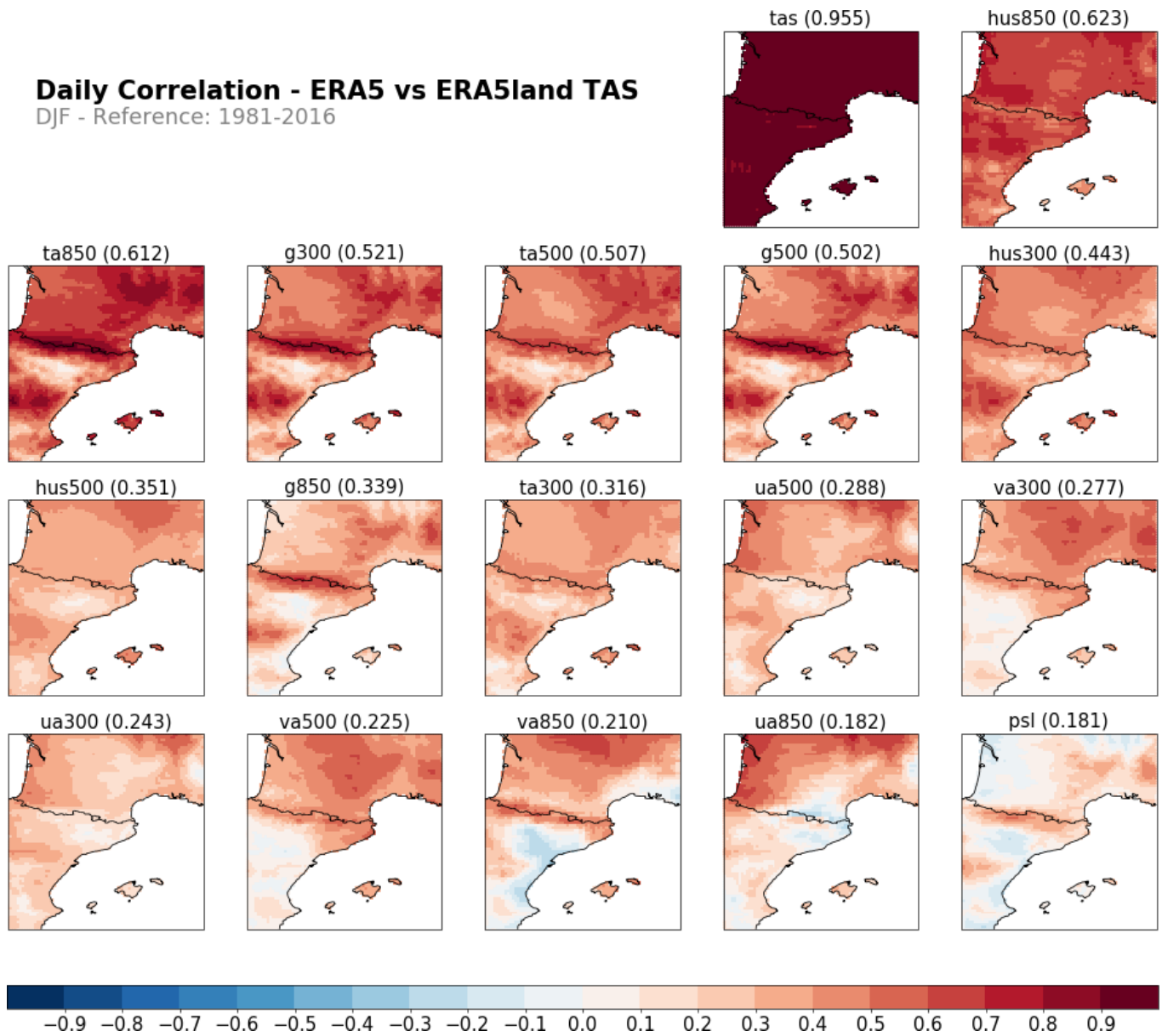


Figure 8: Correlation of the daily time series (DJF) between a set of large-scale predictor variables (ERA5) and local temperature (ERA5-Land). Plots are located in descending order from left to right and top to bottom according to the mean absolute value of all the points in the domain (domain where the training is performed and K-NNs selected), also plotted besides the variable name.

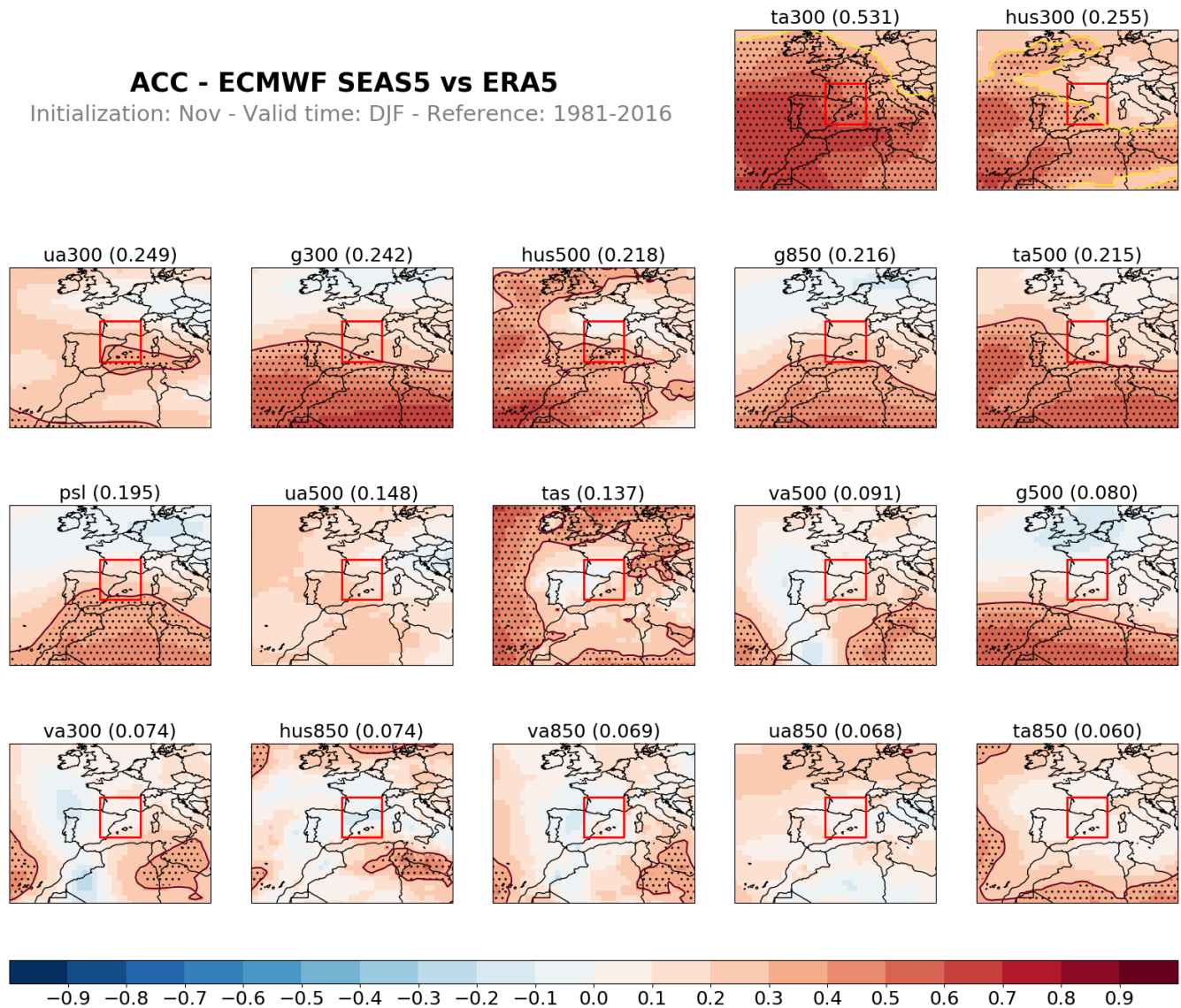


Figure 9: Interannual Anomaly Correlation Coefficient (ACC) between SEAS5 and ERA5 for a set of predictor variables. Plots are located in descending order from left to right and top to bottom according to the mean absolute value of all the points inside the red box (domain where the training is performed and K-NNs selected), also plotted besides the variable name. A ± 0.3 contour and significance (dots) according to a t-test performed over the data are also shown