Assessing seasonal climate predictability using a deep learning application: NN4CAST

Víctor Galván Fraile[1], Belén Rodríguez-Fonseca[1,2], Irene Polo[1], Marta Martín-Rey[1], and María N. Moreno-García[3]

[1]Departamento de Física de la Tierra y Astrofísica, Universidad Complutense de Madrid, Madrid, Spain
[2]Instituto de Geociencias, Consejo Superior de Investigaciones Científicas, Madrid, Spain
[3]Departamento de Informática y Automática, Universidad de Salamanca, Salamanca, Spain

Correspondence: Víctor Galván Fraile (vgalva01@ucm.es); Belén Rodríguez-Fonseca (brfonsec@ucm.es).

Abstract. Seasonal climate predictions are essential for climate services, with changes in tropical sea surface temperature (SST) representing the most influential oceanic drivers. SST anomalies can affect the climate in remote regions through various atmospheric teleconnection mechanisms, and the persistence/evolution of those SST anomalies can give seasonal predictability to atmospheric signals. Dynamical models often struggle with biases and low signal-to-noise ratios, making statistical methods a valuable alternative. Deep learning models are currently providing accurate predictions, mainly in short-range weather forecasts. Nevertheless, the black-box nature of this methodology makes it necessary to ensure its explainability. In this context, we present NN4CAST (Neural Network foreCAST), a Python deep learning pipeline designed to assess seasonal predictability, with built-in tools for evaluating model skill and performing basic spatial diagnostics based on empirical orthogonal functions. Starting from the raw datasets, NN4CAST performs all methodological steps: preprocessing, training and evaluation, enabling researchers to rapidly explore the predictability of a target variable and identify its main potential drivers. This flexible framework allows for the quick testing of predictive skill from different sources of predictability, making it a valuable asset for climate services. Although NN4CAST can use different variables to feed the model, we illustrate its application to reproduce tropical and extratropical teleconnections by training the model with Pacific SST anomalies. We show that NN4CAST can provide skilful predictions across timescales, from modelling variables at lag 0 that capture observed relationships to producing seasonal forecasts at longer leads, both in regions with linear SST-atmosphere coupling (tropics) and in highly non-linear remote regions (such as Europe). Two key examples are the prediction of SST anomalies in the tropical Atlantic region during boreal spring from previous winter SSTs, and the modelling of precipitation anomalies over the European continent in boreal fall from the contemporaneous Pacific SSTs. The former exemplifies a predominantly linear ENSO-Tropical North Atlantic teleconnection, whereas the latter involves a highly non-linear and non-stationary ENSO-Euro-Atlantic teleconnection. Our results demonstrates NN4CAST's potential to determine and quantify the influence of specific potential drivers on a target variable, offering a useful tool for improving climate predictability assessments. NN4CAST enables the attribution of predictions to specific input features, helping to identify the relative importance of different sources of predictability over time and space. In summary, NN4CAST offers a powerful framework to better characterize and understand the complex, non-linear and non-stationary remote climate interactions.

1 Introduction

Seasonal forecasting attempts to provide useful information about the climate that can be expected from 1 up to 12 months. It is crucial for different sectors such as agriculture, water resource management or disaster preparedness. Seasonal predictions are mainly justified by the existence of interactions between the atmosphere and the slow, and predictable, variations in some of the components of the climate system, such as soil moisture, snow cover, stratospheric circulation, ocean heat content or sea surface temperatures (SSTs). Specifically, tropical SSTs are demonstrated to be one of the most important sources of predictability at seasonal timescales due to its characteristic persistence and evolution (Shuila and Kinter III, 2006; Kirtman and Pirani, 2009; Ineson and Scaife, 2009). Earth system models (ESMs) simulate the climate system by solving the mathematical equations that represent the interactions between its components. However, they have errors due to not only the numerical resolution of the system of equations, but also to the uncertainty in the initial and boundary conditions (Hargreaves, 2010). Within the context of seasonal timescales, different challenges, whether occurring independently or in combination, may arise. For example:

1. Seasonal prediction skill relies on the correct representation of both local processes, such as deep convection, and large-scale ones, such as global teleconnections. A misrepresentation of either of them leads to poor performance in certain remote regions, where the interaction of the different signals is non-linear and seasonally dependent (Gleckler et al., 2008; Doblas-Reyes et al., 2013).

2. Oceanic patterns of variability, which may provide seasonal predictability in certain regions such as the Atlantic Niño or the North-Atlantic SSTs, are not well represented by current models (Richter and Tokinaga, 2020; Roberts et al., 2021).

3. Multidecadal ocean variability and the global warming trend alter the global circulation and, consequently, the way in which atmospheric teleconnections (i.e., Rossby waves forced by El Niño) propagate, thereby introducing non-stationarities into the system (López-Parages et al., 2015; Weisheimer et al., 2017). ESMs exhibit mean state biases and struggle to represent interannual-to-decadal variability, failing to reproduce these non-stationarities. Statistical models appear as an alternative to overcome these limitations.

4. The generation of coupled dynamical simulations at these timescales demands substantial computational resources which, in turn, constrains the potential for enhancement in spatial resolution (Doblas-Reyes et al., 2013).

As an alternative to ESMs, statistical methods directly focus on the patterns and relationships between different climate variables with multiple time lags (Wilks, 2011). Another key advantage is their high computational efficiency, as they require significant computing power only during the training phase. Artificial Intelligence (AI) encompasses computational methods that enable machines to learn from data and make predictions (Russell and Norvig, 2016). Within AI, Machine Learning (ML) develops algorithms that identify patterns and generalize to new situations (Samuel, 1959). Deep Learning (DL), a subset of ML using neural networks with multiple hidden layers, excels at capturing complex, high-dimensional relationships without manual feature engineering. Convolutional Neural Networks (CNNs), a widely used DL architecture, learn spatial hierarchies in the data, making them particularly effective for climate applications (LeCun et al., 2015; Rawat and Wang, 2017). The potential

of DL for climate prediction is evident in the growing number of models developed from empirical data rather than explicit physical equations. Recent examples include PanguWeather (Huawei Cloud Group), GraphCast and NeuralGCM (DeepMind and Google) (Bi et al., 2022; Lam et al., 2022; Kochkov et al., 2024), which generate sub-daily forecasts of the Earth system in an autoregressive manner.

The application of deep learning models to seasonal forecasting and the evaluation of teleconnections remains relatively scarce. Other ML techniques such as linear regression, principal component analysis, correlation, and maximum covariance analysis have been widely applied with satisfactory results (Wilks, 2014; Suárez-Moreno and Rodríguez-Fonseca, 2015; Rieger et al., 2021). Despite this, DL offers the potential to overcome current limitations by improving forecast accuracy and enabling analysis of the underlying attribution mechanisms. Existing DL studies have typically targeted individual phenomena, such as El Niño-Southern Oscillation or the Atlantic Niño (Ham et al., 2019; Shin et al., 2022; Bachèlery et al., 2025). However, those approaches produce tailored models for each case, without built-in explainability or generalizability to other teleconnections.

For these reasons, we have developed the Neural Network foreCAST (NN4CAST) application, a Python library designed to facilitate the creation of simple deep learning models for reproducing climate teleconnections driven by different sources of seasonal predictability. NN4CAST provides a flexible framework for non-linear statistical analysis, allowing researchers to efficiently quantify the predictive skill of various sources of predictability. Our approach, in contrast to other ML-based models, facilitates the development of a user-friendly model characterized by a simple architecture and low computational cost. The NN4CAST framework is designed to mitigate the risk of treating DL models as "black boxes", facilitating the interpretation of predictability sources and allowing the assessment of sensitivities to the training period and/or predictor domain. While these capabilities are intrinsic to the framework, the present work focuses on model skill and attribution patterns. This interpretability represents an important added value of NN4CAST compared to other DL seasonal forecasting models (Pan et al., 2022; Watt-Meyer et al., 2024), and is consistent with recent demonstrations that ML weather models trained on reanalysis can yield skilful seasonal predictions while posing interpretability challenges that call for simpler and more transparent experimental setups Kent et al. (2025). It enables the analysis of predictability, as well as the examination of teleconnections, their modulations, the identification of windows of opportunity and the production of attributions of the predictions over certain target regions. NN4CAST is implemented as a Python library intended for use in applications with small or large datasets for seasonal or decadal predictions. Scientific tools are commonly written using low-level compiled languages such as C or C++ (such as the Climate Data Operators, CDO, (Schulzweida et al., 2019)), due to their greater computational efficiency compared to high-level interpreted languages such as Python. However, Python has become the dominant programming language for ML tools, including the libraries used in NN4CAST.

This paper introduces the NN4CAST package, outlining the theoretical foundations of neural networks and detailing its core features. We focus on two well-known climate teleconnections: the Tropical North Atlantic (TNA) SST and European precipitation patterns, which are influenced by Pacific SST anomalies. While these teleconnections are broadly recognized in the literature , the potential of SST to enhance seasonal forecasts has not been fully assessed (Alexander et al., 2002; López-Parages and Rodríguez-Fonseca, 2012). Moreover, the NN4CAST framework extends beyond ocean variables, enabling the investigation of predictability in additional climate components such as soil moisture, snow cover or sea ice. The remainder of the paper is

organized as follows: Section 2 reviews fundamental concepts in DL; Section 3 describes the NN4CAST methodology and implementation; Section 4 presents illustrative applications and evaluates performance; and Section 5 summarizes the main conclusions and outlines directions for future work.

2 Theoretical framework

Seasonal prediction systems typically use a predictor field X initialized at time t_0 to forecast a predictand field Y at $t_0 + \tau$, taking advantage of the persistence of climate anomalies in X. NN4CAST generalizes this setup by enabling the analysis of statistical relationships between X and Y at any temporal lag, from contemporaneous ($\tau = 0$) to predictive configurations ($\tau > 0$). Considering $\tau = 0$ does not imply causality but rather functional relationships, which are crucial for understanding teleconnections. Although numerous statistical methods exist for seasonal prediction, the present work adopts a DL framework. Construction of the neural network entails the selection of hyperparameters, such as network depth, layer widths, activation functions, optimization algorithm and regularization strategies (Géron, 2022). Training is performed by minimizing the Mean Squared Error (MSE) loss, defined as:

$$MSE = \frac{1}{m} \sum_{i=1}^{m} \left(y^{(i)} - \hat{\mathbf{y}}^{(i)} \right)^2 \tag{1}$$

where m is the number of samples in the dataset, $\hat{\mathbf{y}}^{(i)}$ is the predicted value of the predictand variable for the ith instance in the dataset and $y^{(i)}$ is its corresponding real value (i.e., the ground truth) (Wilks, 2011). However, it is possible to modify this objective function to any other differentiable loss function (Cuomo et al., 2022). After training by minimizing the loss function, the model parameters are fixed and used to generate forecasts on the independent test set. Forecast skill is then quantified using the Root Mean Squared Error (RMSE), defined as the square root of the MSE and the Anomaly Correlation Coefficient (ACC), which is given by:

$$ACC = \frac{\sum_{i=1}^{m} y^{(i)} \hat{y}^{(i)}}{\sqrt{\sum_{i=1}^{m} \left(y^{(i)} \right)^2} \sqrt{\sum_{i=1}^{m} \left(\hat{y}^{(i)} \right)^2}} \tag{2}$$

where $\hat{\mathbf{y}}^{(i)}$ is the predicted anomaly of the predictand variable for the ith instance in the dataset, and $y^{(i)}$ is its corresponding observed anomaly, both computed relative to the same climatology (Wilks, 2008).

To achieve a more robust and unbiased evaluation of model performance, NN4CAST implements a cross-validation approach wherein the dataset is systematically partitioned into training and testing subsets across k folds. This iterative process ensures that each sample is utilized for both training and validation, thereby maximizing the use of available data and providing a comprehensive assessment of the model's generalization capabilities. Furthermore, NN4CAST offers the option to perform leave-one-out cross-validation (k=number of samples in the dataset), where each individual sample is used once as a test set while the remaining samples form the training set. This method yields a detailed skill assessment across the entire dataset, facilitating the construction of a full-period hindcast and enhancing the reliability of predictive evaluations (Michaelsen, 1987).

To facilitate a comprehensive analysis of the mechanisms driving the predictions, eXplainable AI (XAI) techniques are employed to assess the relative importance of predictor field features for a given region of the predictand field. XAI aims to

enhance the interpretability of AI methods by identifying the most influential predictor areas in the prediction process. Two main categories of methods exist: sensitivity methods, which assess how changes in a specific predictor affect the output, and attribution methods, which determine the relative contribution of each predictor to the predictand (Guidotti et al., 2018). In this work, we employ one of the most widely used attribution methods, Integrated Gradients (IG), which provides theoretically grounded feature attributions that satisfy desirable axioms such as sensitivity and implementation invariance (Sundararajan et al., 2017). IG is particularly suitable for high-dimensional and complex predictor fields typical of climate data, as it quantifies the contribution of each input relative to a chosen baseline reference vector $\hat{\mathbf{x}}$, for which the model output is zero: $\hat{\mathbf{F}}(\hat{\mathbf{x}}) = 0$. The importance is computed as the product of the distance between the input within the reference point and the average of the gradients at points along the straight-line path from the reference point to the input feature. Specifically, the mathematical expression is given by:

$$R_{i,n} = (x_{i,n} - \hat{\mathbf{x}_i}) \frac{1}{m} \sum_{j=1}^{m} \frac{\partial \hat{\mathbf{F}}}{\partial X_i} \bigg|_{X_i = \hat{\mathbf{x}_i} + \frac{j}{m}(x_{i,n} - \hat{\mathbf{x}_i})} \tag{3}$$

where X_i are the input features, $R_{i,n}$ the relevance of feature at grid point (i) for the model prediction of sample (n), $\hat{\mathbf{F}}$ the function learned by the model and (m) the number of steps in the Riemann approximation (Mamalakis et al., 2022). The attributions computed using this equation provide a quantification of the contribution of each input feature to the predicted output for a given target region. By integrating gradients along the path from a baseline input to the actual predictor field, we can identify the most influential regions and quantify their relative importance. These attributions reflect statistical associations captured by the model and should not be interpreted as evidence of direct causal relationships. This approach allows the model to remain entirely data-driven, while providing interpretable information at the feature level on the factors that determine seasonal predictability, in line with XAI principles. The next section describes how this theoretical framework is implemented in the NN4CAST tool.

3 NN4CAST methodology & implementation

Building on the key elements of the DL methodology outlined above, the NN4CAST library integrates these approaches into a unified framework. The whole procedural workflow is depicted in Fig. 1:

· Preprocessing of the datasets according to user targets, including: data loading, selection of the region and season of interest, anomaly computation, and trend removal (I in Fig. 1).

· Construction of a deep neural network for seasonal prediction, with training performed through the minimization of the MSE (II in Fig. 1).

· Application of regularization techniques to mitigate overfitting (II in Fig. 1).

· Model performance evaluation using cross-validation strategies, employing different skill metrics (RMSE and ACC, II in Fig. 1).
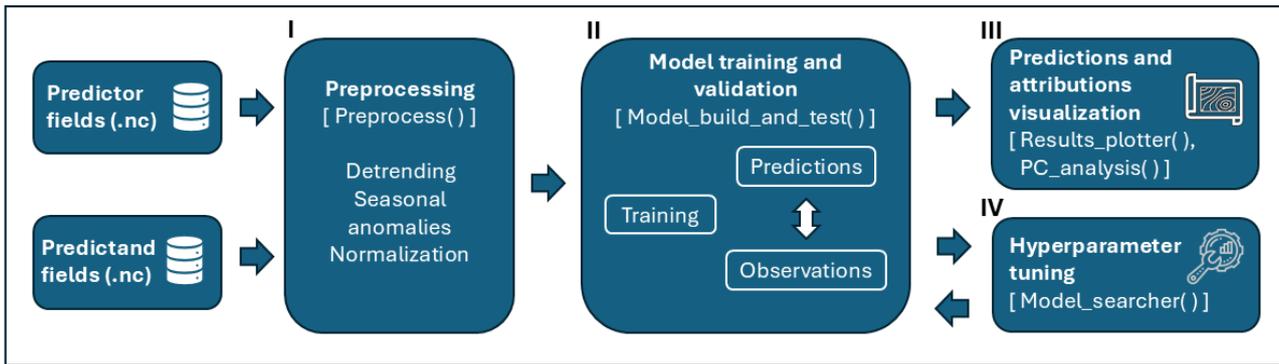
Figure 1. Flowchart illustrating the methodology and application workflow of the NN4CAST library, designed for processing monthly data and assessing seasonal climate predictability. The Python function names corresponding to each step are shown in brackets.

· Attribution analysis within the XAI framework to identify the contribution of different regions of the predictor field to the model predictions (III in Fig. 1).

· Empirical Orthogonal Function (EOF) analysis of model outputs and observational data to compare dominant modes of variability and assess the physical consistency of predictions (III in Fig. 1).

· Optimization of model hyperparameters to improve generalization and predictive performance (IV in Fig. 1).

Each of the above mentioned steps is computed by the application of different Python functions designed within NN4CAST (Fig. 1). The application begins by loading predictor and predictand datasets and defining model hyperparameters. A standardized preprocessing pipeline is then applied via the `Preprocess()` function. Next, the model is built, trained, and tested using `Model_build_and_test()`, which integrates attribution routines to compute feature importance maps for each forecast. Once training is complete, `Results_plotter()` visualizes both the individual predictions and their associated attribution maps, making the physical drivers of skill explicit. Optionally, `Model_searcher()` can be used to optimize performance through hyperparameter tuning, and `PC_analysis()` enables EOF analysis of model outputs versus observations to assess dominant spatial modes of variability. Finally, the entire workflow can be configured via the main module `nn4cast.predefined_classes`.

The principal distinguishing feature of NN4CAST lies in its design, which specifically targets the assessment of teleconnection predictability, enabling experiments to address common challenges in seasonal climate forecasting. While some of these challenges could be partially addressed by simpler models, NN4CAST leverages deep learning to capture complex, nonlinear relationships and spatial interactions that are often missed by linear approaches. The main challenges in seasonal climate forecasting, discussed in the introduction, are tackled by NN4CAST through the following strategies:

1. Representing teleconnection drivers at interannual timescales. The models created by NN4CAST can capture nonlinear and spatially distributed interactions among predictors across different regions, enabling a more accurate assessment of each region contribution to the predictand.

2. Non-stationarity of climate relationships. The models created by NN4CAST can be trained over different periods, enabling the analysis of how predictor-predictand relationships evolve over time.

3. Computational efficiency. The models created by NN4CAST are optimized to perform simulations within minutes on a standard computer, despite modelling high-dimensional and complex predictor fields.

4. Interpretability of DL models. The models created by NN4CAST provide feature attributions that reveal the main drivers and underlying physical mechanisms of predictions, combining the flexibility of DL with interpretability.

3.1 Configuration of hyperparameters and Preprocessing

NN4CAST employs datasets stored in the netCDF4 format (Rew et al., 2006), a widely used standard within Earth data science. These datasets are structured as space-time matrices and require three coordinates: time, latitude, and longitude. As NN4CAST is designed for seasonal timescales, the time coordinate must be defined at a monthly resolution. The hyperparameters required by the application are summarized in Table 1, including a brief description of their functionality. These hyperparameters are stored in a dictionary that can be saved as a YAML file to document the experiment setup for easy retrieval and reproducibility. The next step involves preprocessing the predictor and predictand data according to the operations specified in the hyperparameter dictionary, such as regridding, computing seasonal means, and detrending. Detrending is performed using a backward moving average (BMA) algorithm, which computes the running mean of previous years using a sliding window and subtracts it from subsequent values (Raffalovich, 1994; Alvarez-Ramirez et al., 2005). This approach ensures that future information is not introduced during preprocessing. For a complete description of the code implementation, including the creation of the hyperparameter dictionary and the preprocessing workflow, the reader is referred to the Supplementary Material (Listings S1-S3).

Hyperparameter name	Description
$path_x, path_y$	Path to the X (predictor) and Y (predictand) datasets files.
$time\_limits$	Define the time limits (in years) to select from the original datasets. They should be in the format ['yyyy', 'yyyy'], where the first year is the start and the second year is the end of the period, separated by commas. For example, ['1975','2015'].
$jump\_year$	Specify whether the predictor (X) and predictand (Y) fields are drawn from the same calendar year or from different years. A value of 0 (default) indicates same-year data; any positive integer denotes the lead time in years between X and Y. For example, if X corresponds to September of year N and Y to February of year N+1, setting $jump\_year = 1$ relates the predictor and predictand across consecutive years.
$months_x, months_y$	Select the months of X and Y, knowing that January=1 and December=12.
$name_x, name_y$	Specify the variable names for the predictor and predictand fields, as defined in the original netcdf files.
$units_x, units_y$	Set the units of the input and output fields to fill the labels in the plot.

$region\_predictor$	Set the name of the predictor region to be included as information for the figures.
$months\_skip\_x, months\_skip\_y$	Define the months to skip. This is necessary if several months are considered to analyse seasons involving changes of the year. For example, if X samples are from Dec-Jan-Feb and the data spans from 1950 to 2019, it is necessary to set $months\_skip\_x$ = ['1950-01', '1950-02', '2019-12'], to delete the Jan-Feb from the first year and the Dec from the last one, in order to create seasonal means that makes climatic sense.
$train\_years, validation\_years,$ $test\_years$ $mean\_seasonal\_method\_x,$ $mean\_seasonal\_method\_y$	Define the periods (initial year, final year) for training, validating and testing the model based on X years, using the same format as for $time\_limits$. Define if computing the seasonal means (=True) or aggregates (=False).
$lat\_lims\_x, lon\_lims\_x,$ lat_lims_y, lon_lims_y	Define the latitude-longitude regions for the predictor (X) and predictant (Y) fields. For the latitudes, it can be selected in any order (smaller first or last), whereas, for the longitudes, it is needed to put the smaller first, and either in the format: -180-(+180) or 0-360.
$scale\_x, scale\_y$	Define if scaling the data, if not, set as 1 (default=1). For example, if the pressure data is given in Pa and you want to have the data in hPa, then, a scale of 100 need to be applied.
$regrid\_degree\_x,$ $regrid\_degree\_y$	Define the interpolation degree for regridding the data. Values greater than 0 apply regridding, while 0 means no regridding (default=0).
$detrend\_x, detrend\_y$	Define if performing a backward moving average detrend of the data (default=False). If True, it is needed to define the window size to compute the detrend iteratively.
$detrend\_x\_window,$ $detrend\_y\_window$	Define the size of the past window to compute the averages to do the detrend (default=10).
$layer\_sizes$	Define the number of neurons per layer of the model as a list. For example, [1024, 256, 64, 256, 1024].
$activations$	Define the activation function of each layer of the model as a list. For example, [tf.keras.activations.elu, tf.keras.activations.elu, tf.keras.activations.elu].
$kernel\_regularizer$	Define whether using or not a regularizer. The options are: $l\_1, l\_2, l1\_l2$ or $None$ (default=['$l1\_l2$']).
$learning\_rate$	Set the learning rate of model hyperparameters update (default=0.0001).
$epochs$	Set the number of epochs to train the model, knowing that there is an early stopping when no increase is reached in the validation loss on the last 25 epochs (default=2500).
num_conv_layers	Set the number of convolutional layers to apply (default=0). If at least 1 convolutional layer is applied, it is added at the input of the network, before the dense (fully connected) layers. After the convolutional processing, the resulting features are flattened and concatenated with the dense layers, allowing the network to integrate both convolutional and fully connected representations. In this case, you can also define the $number\_filters$, $kernel\_size$, and $pooling\_size$.
$use\_batch\_norm,$ $use\_initializer$	Set to True to perform batch normalization and Kaiming He initialization (He et al., 2015) of the parameters in each layer of the model. (default=False).
$use\_dropout, dropout\_rates$	Set to True to perform dropout, giving also the dropout rates as a list. (default=False).
$use\_init\_skip\_connections,$ $use\_inter\_skip\_connections$	Set to True to use initial or intermediate skip connections as in the U-Net architecture (Ronneberger et al., 2015) (default=False).
p_value	Define the p_value used to establish the confidence interval through the T-test method for assessing statistically significant values of the model skill.
$outputs\_path$	Select the directory where the plots and datasets will be saved.

Table 1. Table listing the names and detailed descriptions of the hyperparameters used in the NN4CAST application, which govern the model configuration and training process. Default values have been optimized for various climate variability studies using the built-in hyperparameter optimization function `Model-searcher`. These defaults provide reliable performance, although some hyperparameters can be adjusted depending on the specific dataset or application.
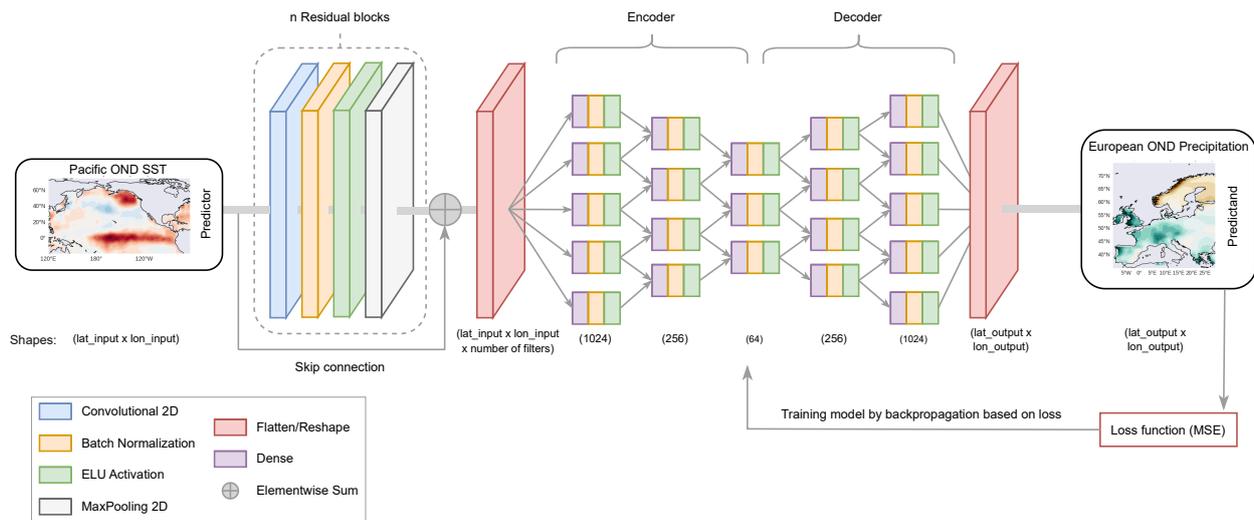
Figure 2. Schematic of the NN4CAST model architecture used in this study to analyse the predictability of the Pacific-European precipitation teleconnection (4.2). The network consists of an initial block of convolutional layers (if the hyperparameter num_conv_layers is set to a value different from 0), followed by several fully connected (dense) layers, whose number and size are defined by the hyperparameter $layer\_sizes$ (see Tab. 1 for a detailed explanation of the architectural parameters).

3.2 Layer-by-layer specification of the default architecture

To ensure full reproducibility, we provide a layer-by-layer description of the default architecture used in the Pacific–European teleconnection experiment (Sect. 4.2), corresponding to the hyperparameter configuration in Table 3. For this configuration, `num_conv_layers = 0` and no skip connections are used; thus, the model reduces to a fully connected encoder–decoder network. If convolutional layers were included, they would use two-dimensional convolutions with unit stride and same padding, preserving the spatial dimensions. For an input tensor of size (H, W), the output would retain the same spatial resolution, (H, W), while the channel dimension would become equal to the number of filters F, yielding a tensor of shape (H, W, F). After flattening, this corresponds to a vector of size $H \times W \times F$. For each hidden dense layer, the following sequence of operations is applied: Dense $\rightarrow$ Batch Normalization $\rightarrow$ ELU activation. Batch normalization, kernel regularizer and dropout are included only when `use_batch_norm=True`, `use_initializer=True` and `use_dropout=True`, respectively. When enabled, kernel regularization and dropout are applied immediately after the flattening layer. In the present configuration, a dropout rate of 0.1 is used. The output layer employs a linear activation function and does not include dropout. The resulting output vector is reshaped to the predictand field size. Model parameters are optimized by minimizing the mean squared error (MSE) using the Adam optimizer with a learning rate of 10^{-4}. A schematic representation of the architecture is shown in Figure 2.

3.3 Model Cross-Validation and Explainable AI

To obtain robust and objective estimates of model performance, NN4CAST includes a cross-validation routine via the `Model_build_and_test()` function. While a standard train-validation-test split (e.g., 70%-10%-20%) is possible, cross-validation is generally recommended as it maximizes the use of available data and reduces sensitivity to the arbitrary selection of training periods. By setting the `n_cv_folds` parameter, the dataset is partitioned into k folds: in each iteration, the model is trained on $k-1$ folds and tested on the remaining one, cycling through all folds. A random 10% of each training fold is reserved for early stopping. When k equals the number of samples, this defaults to leave-one-out cross-validation, producing a complete hindcast over the full period.

A core objective of NN4CAST is to facilitate the generation of insightful attributions between predictions. To compute these attributions for a specific target region, the user can enable the attribution calculation and define the region of interest as a list of latitude and longitude ranges. The model then applies the Integrated Gradients methodology across the entire hindcast, leveraging the cross-validation outputs.

3.4 Hyperparameter Optimization

NN4CAST also provides a functionality to optimize the model hyperparameters for improved performance. The user first defines a search space for relevant hyperparameters,and then applies the `Model_searcher()` function with a specified maximum number of trials. The evaluation follows a cross-validation scheme to ensure robustness and objectivity. For full details of the cross-validation workflow, hyperparameter search setup, and code implementation, see Supplementary Material (Listings S3-S4).

4 Applications of NN4CAST

4.1 Modelling Pacific-North Tropical Atlantic SST Connection

In this section, we present a case study to illustrate the applicability of the NN4CAST framework for modelling a well-documented atmospheric teleconnection. Specifically, we focus on the relationship between SST anomalies in the tropical Pacific during boreal winter (December-January-February; DJF) and SST anomalies in the tropical North Atlantic (TNA) during the subsequent spring (March-April-May; MAM). This teleconnection has been extensively studied in the literature (Enfield and Mayer, 1997; Alexander et al., 2002; Lee et al., 2008; García-Serrano et al., 2017), making it an ideal test case to demonstrate the capabilities of NN4CAST. It is well established that El Niño events during boreal winter are typically associated with warming in the TNA region through a process commonly referred to as the atmospheric bridge between the Pacific and Atlantic Oceans (Alexander et al., 2002). Several mechanisms have been proposed to explain this teleconnection. One hypothesis suggests that the weakening of the subtropical high over the North Atlantic is driven by atmospheric Rossby waves originating in the tropical Pacific and propagating via the Pacific-North American (PNA) pattern. (Enfield and Mayer, 1997). Another mechanism suggests that changes in the Pacific zonal atmospheric circulation (Walker cell) may influence

Parameter	Value	Parameter	Value
Input/Output, Temporal and Spatial Settings			
region_predictor	Tropical Pacific	time_lims	1900-2019
lat_lims_x	30, -30	lon_lims_x	120, 290
lat_lims_y	40, -10	lon_lims_y	-80, 20
Preprocessing			
detrend_x/y	True/True	detrend_x/y_window	30/30
mean_seasonal_method_x/y	True/True	regrid_degree_x/y	2/2
months_x	12, 1, 2	months_y	3, 4, 5
Model Architecture			
layer_sizes	1024, 256, 64, 256, 1024	activations	ELU
num_conv_layers	0	dropout_rates	0.1
use_dropout	True	use_batch_norm	True
use_initializer	True	kernel_regularizer	l2
use_init_skip_connections	False	use_inter_skip_connections	False
Training Settings			
learning_rate	0.0001	epochs	2500
p_value	0.1	jump_year	0
cross_validation	True	CV strategy	Leave-One-Out

Table 2. Table summarizing the selected model hyperparameters and preprocessing settings used in the simulation of DJF Tropical Pacific - MAM Tropical North Atlantic SSTs.

convection in the Atlantic, which in turn affects the Atlantic meridional atmospheric circulation (Hadley cell) (Klein et al.,
240 1999). Both mechanisms ultimately lead to a weakening of the subtropical high, resulting in reduced trade winds and enhanced
SST warming in the TNA. Additional theories in the literature point to upper-tropospheric equatorial responses to Pacific SST
anomalies, which can trigger eastward-propagating Kelvin waves (Chang et al., 2001), and a remote Gill-type response to
ENSO-related changes in the Pacific Walker circulation, whose surface signature affects the winds and contributes to warming
of the TNA (García-Serrano et al., 2017). The teleconnection between ENSO and TNA SST anomalies depends on the type
245 and persistence of ENSO events, which determine the strength and duration of the atmospheric bridge. The North Atlantic
Oscillation (NAO), typically following modifications of the PNA pattern, can modulate the TNA response: its negative phase is
associated with a weakened subtropical high, thereby enhancing the influence of El Niño on TNA SSTs (Lee et al., 2008; Czaja
et al., 2002; Wu et al., 2020).

Using the NN4CAST framework, a tailored model can be efficiently developed to simulate the Pacific-Atlantic teleconnection,
250 assess its predictive skill across various events, and explore attribution patterns for selected case studies. This allows investigation
of whether the predicted TNA is mechanistically related to changes in the teleconnection. These patterns can then be compared

with mechanisms proposed in the existing literature. To analyse this teleconnection using NN4CAST, we utilise the HadISST dataset as the source of sea surface temperature (SST) data for both predictor and predictand fields (Rayner et al., 2003). Additionally, surface winds (U10, V10) and geopotential at 200 hPa (Z200) from ERA-20C reanalysis is used for dynamical analysis (Poli et al., 2016). The predictor region corresponds to the tropical Pacific basin [30°S-30°N; 120°E-70°W]. This region captures the core variability associated with ENSO phenomena. The predictand region corresponds to the TNA [10°S-40°N; 80°W-20°E], encompassing the key area where the SST anomalies linked to this teleconnection typically emerge during boreal spring. The parameters and hyperparameters used are the ones shown in Table 2. First, the dictionary with the details of the simulation is created and saved in the outputs directory. Then, the preprocessing of the datasets is done by applying as optional main arguments: regriding the data to reduce the computational cost of the simulation and detrending to remove the signal associated with the anthropogenic warming trend. Subsequently, the model is initialised using the hyperparameters specified in Table 2, adopting an encoder-decoder architecture (Hinton and Salakhutdinov, 2006; Goodfellow et al., 2016). This architecture first compresses the high-dimensional predictor field into a low-dimensional latent representation and then reconstructs the target field from that embedding. The bottleneck design facilitates the efficient extraction and interpretation of the most relevant predictive features while mitigating overfitting in high-dimensional spaces. Similar approaches have also been applied in seasonal prediction tasks. For example, (Ibebuchi and Richman, 2024) combined autoencoders with LSTM networks to forecast ENSO, showing that compact latent representations can capture physically meaningful patterns and improve predictive skill. Although our implementation is based on a fully connected architecture, it follows the same principle of dimensionality reduction and reconstruction, providing a balance between model simplicity and interpretability.

To obtain a robust measure of model performance, we apply a cross-validation approach. This is achieved by enabling the cross-validation argument and specifying the number of folds into which the dataset will be partitioned. Specifically, a leave-one-out cross-validation scheme with 120 folds is employed. The metrics of model performance across the entire hindcast period are summarized in the output figure of the function (Fig. 3). High skill is shown in the whole tropical Atlantic region, with maximum values (up to 0.9) over the western side of the basin (Fig. 3a). Notably, the skill in the TNA region remains positive and ranging from 0.4 to 0.8 in most of the years, with improved performance during certain decades, as during the 1980s, where each yearly value represents the spatial anomaly correlation computed over the full domain (Fig. 3b). In terms of the RMSE (Fig. 3c), three regions stand out where model errors are more pronounced: the Atlantic Niño, the Gulf Stream, and the SMSCU system. These areas coincide with regions exhibiting the highest SST variability. Interestingly, the model achieves a high level of skill using only tropical Pacific information, allowing the impact of boreal winter ENSO conditions on the subsequent spring tropical Atlantic state to be quantified. Furthermore, to identify the regions from which the model extracts information to generate the TNA signal, attributions for two regions are computed using the Integrated Gradients method, as outlined in Section 2: one representing the western TNA (WTNA) region [10°N-20°N; 50°W-70°W], where the model shows higher predictive skill (Fig. 3a), and the other representing the southern Mauritanian-Senegalese Coastal Upwelling (SMSCU) system [10°N-20°N; 25°W-15°W], where prediction errors are larger (Fig. 3b). These regions were also chosen based on previous research on the impact of ENSO-related teleconnections in both areas (García-Serrano et al., 2017; Wade et al., 2023).
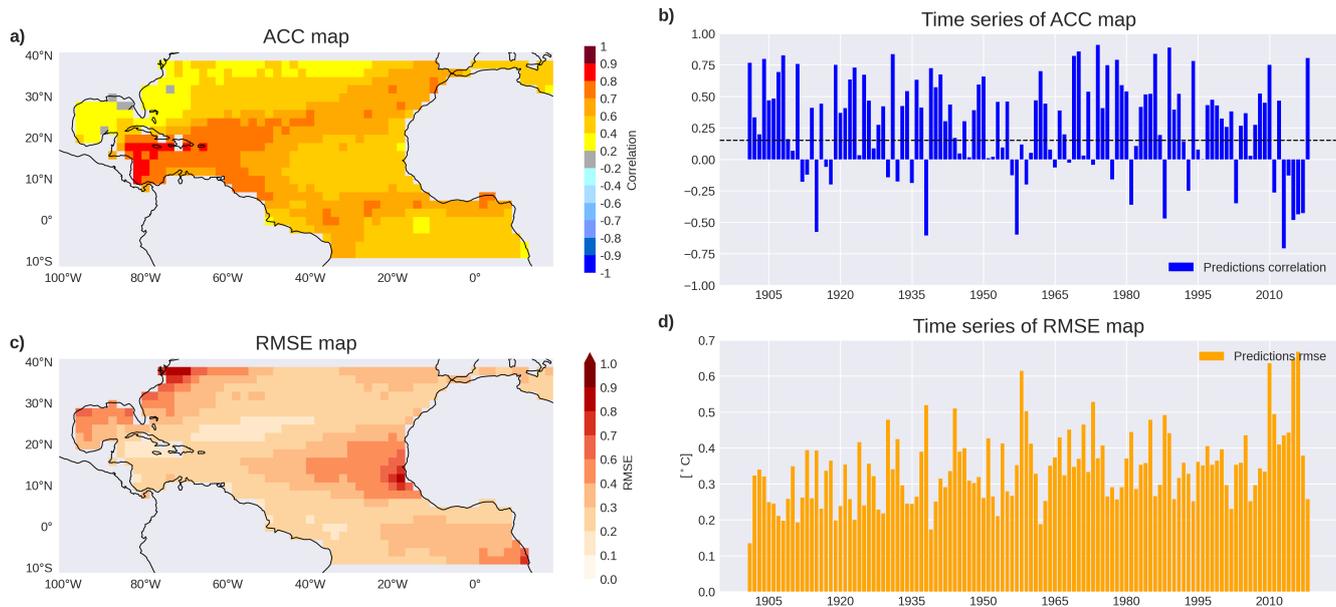
Figure 3. Predictability of tropical North Atlantic SST variability from tropical Pacific anomalies. Panels show model performance over 1901-2019 using leave-one-out cross-validation to predict MAM TNA SST anomalies from DJF tropical Pacific SST anomalies. Predictions are compared with observed MAM SST anomalies. Specifically: (a) Spatial anomaly correlation coefficient (ACC) map, computed at each grid point as the temporal correlation (1901-2019) between predicted and observed anomalies; (b) Yearly spatial ACC time series, where each value represents the spatial anomaly (pattern) correlation between predicted and observed anomaly fields over the full TNA domain for a given year; (c) Spatial RMSE map, computed analogously to (a); and (d) Yearly spatial RMSE time series, computed analogously to (b). Shading in (a) indicates statistically significant correlations at the 95% level (one-tailed t-test). The dashed line in (b) marks the corresponding significance threshold.

To better understand the model behaviour, we constructed composites based on the WTNA and SMSCU indices, selecting events where the respective indices exceeded ±0.5 standard deviations. Figure 4 shows composites of different fields for the positive events of both indices. Panels (a) and (c) display the predicted Atlantic SST for positive WTNA and SMSCU events, respectively, including surface wind anomalies in MAM to highlight local dynamical changes. The WTNA composite shows weaker wind anomalies close to the African coast, consistent with the weaker predicted SST signal, whereas the SMSCU positive composite reveal a local strengthening of southwesterly winds along the Senegalese coast, which can contribute to a reduction of the coastal upwelling and strong coastal SST warming. Attribution maps highlight that the central Pacific is a major contributor for both WTNA and SMSCU, specifically around 170ºE-150ºW, with additional positive contributions from the eastern part of the Pacific basin as well as negative contributions from a region located around 130°W-110°W (Fig. 4b-d). Notice that these spatial patterns should be interpreted as reflecting statistical associations within the learning framework, capturing co-varying large-scale climate signals rather than direct causal influences from individual regions.
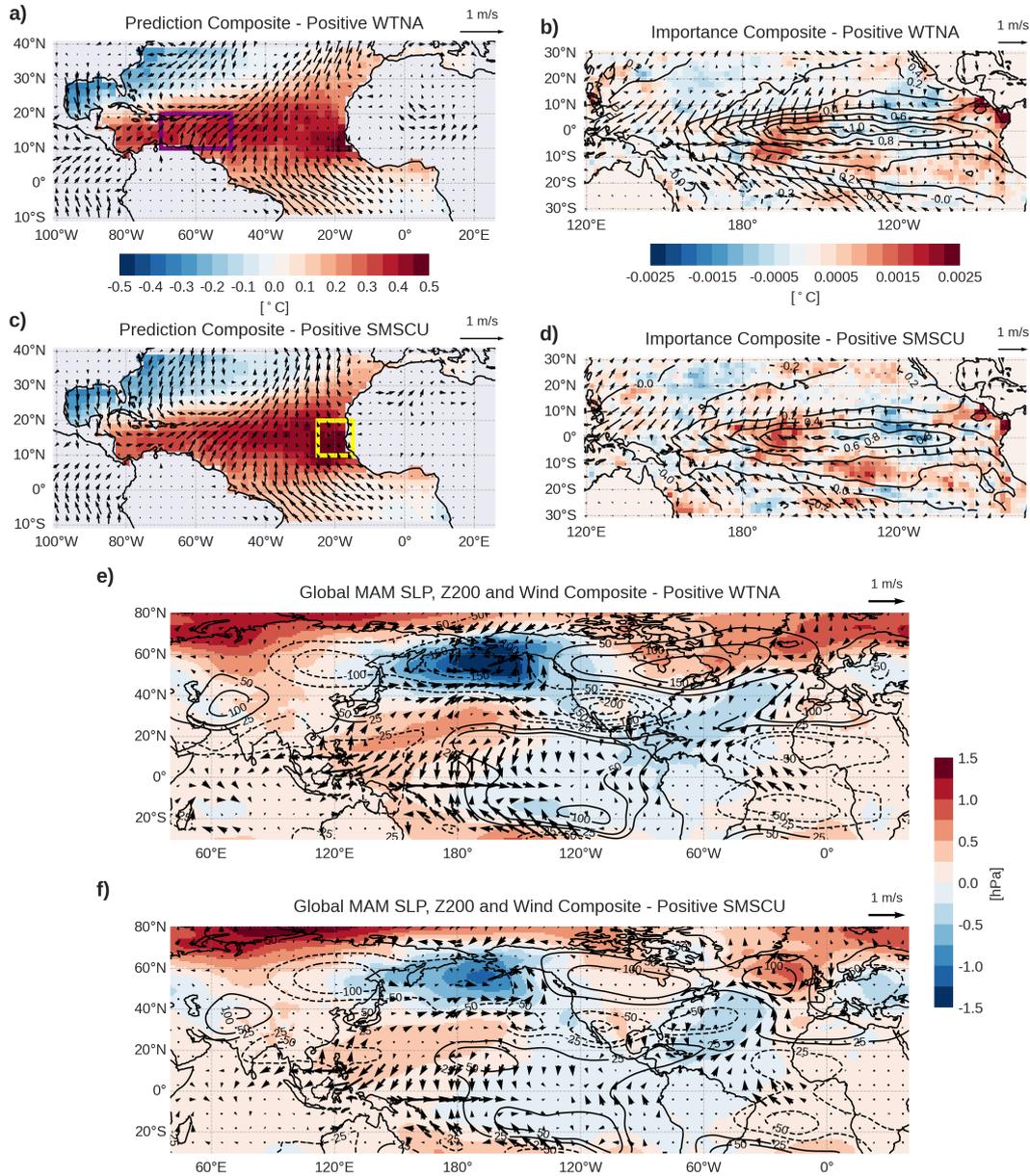
Figure 4. Composites of model anomalous SST predictions, predictor fields, and attribution maps for positive predicted WTNA and SMSCU, based on 28 and 26 events, respectively. Panels (a) and (c) show the predicted mean SST anomalies in the Atlantic during MAM together with surface wind anomalies indicated by arrows. Panels (b) and (d) show the attribution maps over the predictor fields with SST in contours and surface winds in arrows. Panels (e) and (f) display global composites of MAM anomalies in sea level pressure (shading), 200 hPa geopotential (contours), and surface winds for positive WTNA and SMSCU events. Attribution maps indicate the relative contribution of each grid point in the predictor field to the forecasted value in the target region, with the sum of the values within each map matching the predicted anomaly in the corresponding index region (i.e., the sum of all values in panel (c) matches the WTNA anomaly within the purple box in panel (a)).

14

The region of positive attributions is coherent with physical mechanisms linking both basins. Indeed, WTNA predicted anomalies corresponds to events in which the central Pacific presents a pronounced wind convergence, consistent with a Gill-type atmospheric response characteristic of a central Pacific warming (Fig 4e) (Gill, 1980). This tropical response initiates a broader extratropical wave propagation towards the Atlantic, which is related with a negative NAO-like pattern that weakens the trade winds over the TNA region, particularly for WTNA events (Horel and Wallace, 1981; Czaja et al., 2002). A secondary Gill-type response over the equatorial Atlantic, associated with the anomalous upper-level convergence induced by the modification of the Walker cell, further contributes to the weakening of the trade winds on its western flank and to reduced upwelling. In the case of warm phases of the SMSCU, this secondary Gill response is more regional, and the associated warming could be linked to anomalous negative Ekman pumping, which reduces the upwelling (Calvo-Miguélez et al., 2025). Model predictions and their attributions for two specific samples, 1986 and 1997, as well as composites for the negative phases of WTNA and SMSCU, and Niño3.4-based composites, are provided in the Supplementary Material (Figs. S1-S5).The Supplementary Material also includes regression maps of the WTNA and SMSCU indices onto Pacific SSTs, allowing for comparison with the model attributions.

4.2   Modelling North and Tropical Pacific - European Precipitation Teleconnection

In this section, we focus on the relationship at the seasonal scale between north and tropical Pacific SST anomalies and European precipitation during boreal autumn (October-November-December; OND), a challenging target variable for dynamical models (Johnson et al., 2019). In the North Atlantic, the dominant mode of atmospheric interannual to decadal variability is the NAO, which modulates the pressure gradient between the Icelandic Low and the Azores High (Rogers, 1997; Trigo et al., 2002). This gradient, in turn, determines the storm tracks over Europe and consequently influences precipitation. Although the NAO signal largely arises from internal variability, external factors such as ENSO can modulate its centers of action via tropospheric and stratospheric pathways (Rodwell et al., 1999; Rodríguez-Fonseca et al., 2016). The interplay with other climate modes, including the Atlantic Multidecadal Oscillation (AMO) and the Pacific Decadal Oscillation (PDO), contributes to the non-stationarity of this teleconnection and, consequently, to the varying influence of ENSO on European precipitation over time (López-Parages and Rodríguez-Fonseca, 2012).

For this analysis, SST data from the HadISST dataset were used as the predictor field (Rayner et al., 2003), while precipitation data from the Climatic Research Unit gridded Time Series (CRU TS) dataset served as the predictand (Harris et al., 2020). Additionally, surface winds (U10, V10) and geopotential and zonal winds at 200 hPa (Z200, U200) from ERA-20C reanalysis is used for dynamical analysis (Poli et al., 2016). The predictor region covers the Pacific basin [20°S-75°N; 120°E-70°W] to capture the ENSO and extratropical Pacific related variability, while the predictand region is defined by the European continent [35°N-75°N; 10°W-30°E]. The hyperparameters and other modelling parameters are detailed in Table 3. Unlike the previous case study, precipitation was aggregated from monthly data (by setting the parameter $mean\_seasonal\_method\_y = False$).

After performing a leave-one-out cross-validation, the results are summarized in Figure 5. The ACC map indicates that only northern Scandinavia, northern Spain and eastern central Europe exhibit significant positive skill (Fig. 5a). The RMSE map indicates higher errors in regions with greater precipitation variability, such as Norway, the United Kingdom and northwestern Spain (Fig. 5c). At first glance, these results might suggest that either the model fails to capture the teleconnection or that the

Parameter	Value	Parameter	Value
Input/Output, Temporal and Spatial Settings			
region_predictor	Pacific	time_lims	1901-2019
lat_lims_x	75, -20	lon_lims_x	120, 300
lat_lims_y	75, 35	lon_lims_y	-10, 30
Preprocessing			
detrend_x/y	True/True	detrend_x/y_window	30/30
mean_seasonal_method_x/y	True/False	scale_x/y	1/1
regrid_degree_x/y	2/1	months_x	10, 11, 12
months_y	10, 11, 12	months_skip_x/y	None/None
Model Architecture			
layer_sizes	1024, 256, 64, 256, 1024	activations	ELU
num_conv_layers	0	dropout_rates	0.1
use_dropout	True	use_batch_norm	True
use_initializer	True	kernel_regularizer	l2
use_init_skip_connections	False	use_inter_skip_connections	False
Training Settings			
learning_rate	0.0001	epochs	2500
p_value	0.1	jump_year	0

Table 3. Table summarizing the model hyperparameters and preprocessing settings used in the simulation of Pacific SSTs and European precipitation in OND.

Pacific SST and European precipitation link is inherently non-stationary, as suggested by (López-Parages and Rodríguez-Fonseca, 2012). The temporal evolution of the ACC and RMSE (Fig. 5b-d), where each yearly value represents the ACC and RMSE computed over the full domain, supports the latter hypothesis; there are periods such as the 1950s-1970s when the model displays considerable skill (ACC > 0.4), contrasted with periods like the 2000s where skill is minimal (ACC ≈ 0).

335   To gain further insight into the model predictions and to understand the type of European precipitation variability that NN4CAST is capturing, we conduct additional analyses. Concretely, we employ the built-in `PC-analysis` functionality of NN4CAST to extract the EOFs. The leading mode of variability of the OND precipitation fields, explaining 30% and 44% of the total variance in the observational and model prediction datasets respectively, exhibits a coherent spatial pattern, indicating that the model reproduces the variability with remarkable accuracy, although overestimating the relative importance of this leading

340 mode (Fig. 6a-b). In particular, the model effectively captures the core features of this mode, especially in regions where this mode of variability has the higher impacts, such as Scandinavia, the United Kingdom, western Europe, and the Balkans. While a slight underestimation in the overall amplitude is evident, the spatial correlation between the simulated and observed patterns
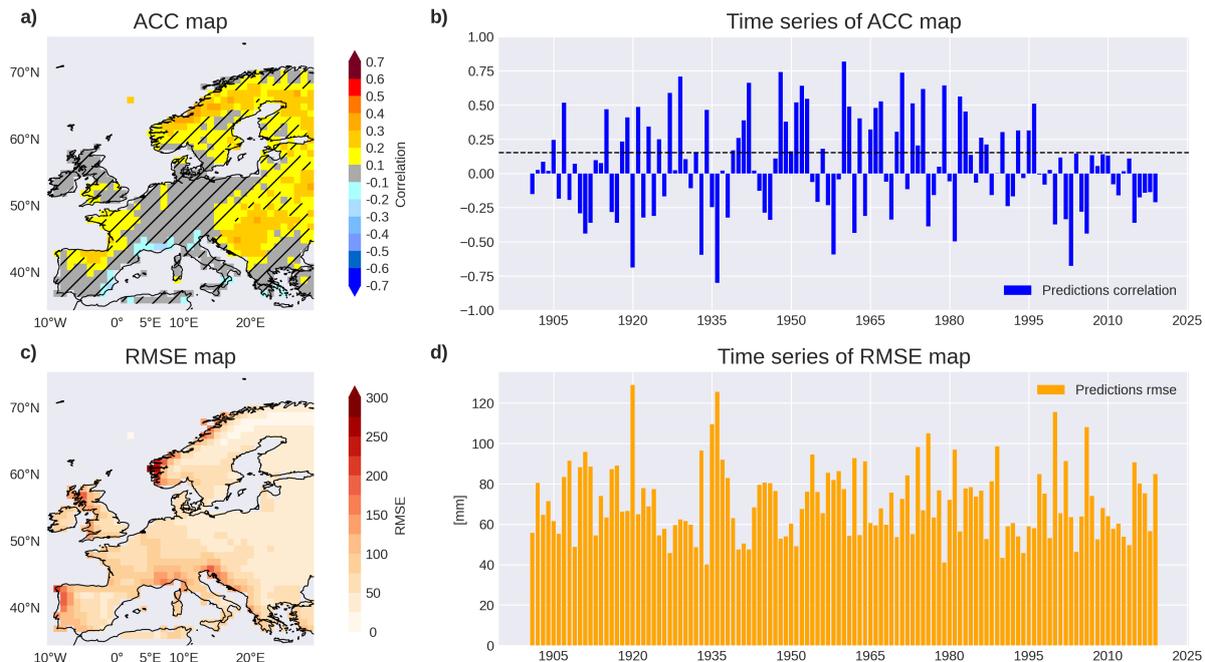
Figure 5. Predictability of European precipitation variability from north and tropical Pacific SST anomalies. Panels show model performance 1901-2019 using leave-one-out cross-validation to predict OND European precipitation anomalies from OND Pacific SST anomalies. Specifically: (a) Spatial anomaly correlation coefficient (ACC) map, computed at each grid point as the temporal correlation (1901-2019) between predicted and observed anomalies; (b) Yearly spatial ACC time series, where each value represents the spatial anomaly (pattern) correlation between predicted and observed anomaly fields over the full European domain for a given year; (c) Spatial RMSE map, computed analogously to (a); and (d) Yearly spatial RMSE time series, computed analogously to (b). Shading in (a) indicates statistically significant correlations at the 95% level (one-tailed t-test). The dashed line in (b) marks the corresponding significance threshold.

reaches r = 0.98, indicating that the model skilfully represents the structure and geographical distribution of the dominant precipitation variability during this season.

345   To better understand how this variability evolves over time, the associated PC time series were first temporally filtered to retain only their high-frequency components by applying a Butterworth filter with a 7-year cut-off frequency, following the methodology of other works (López-Parages et al., 2015). Twenty-year centered moving-window correlations between the observed and predicted PCs (Fig. 6c) reveal that model skill exhibits limited skill during 1910-1940, improved skill during 1940-1969 (r≈0.4-0.7), and modest skill thereafter 1970-2010 (r≈0.2-0.5). To identify the potential climate driver behind these

350   fluctuations, we then calculated the moving-window correlations between each high-frequency PC and the Niño 3.4 index (Fig. 6d). During the 1940-1969 period of higher skill, the PC-Niño 3.4 correlation is negative, indicating that La Niña events coincide with increased precipitation over western Europe. In contrast, another period of significant correlations shows positive relationships between the leading PC and El Niño. The change in sign across periods indicates how different teleconnections
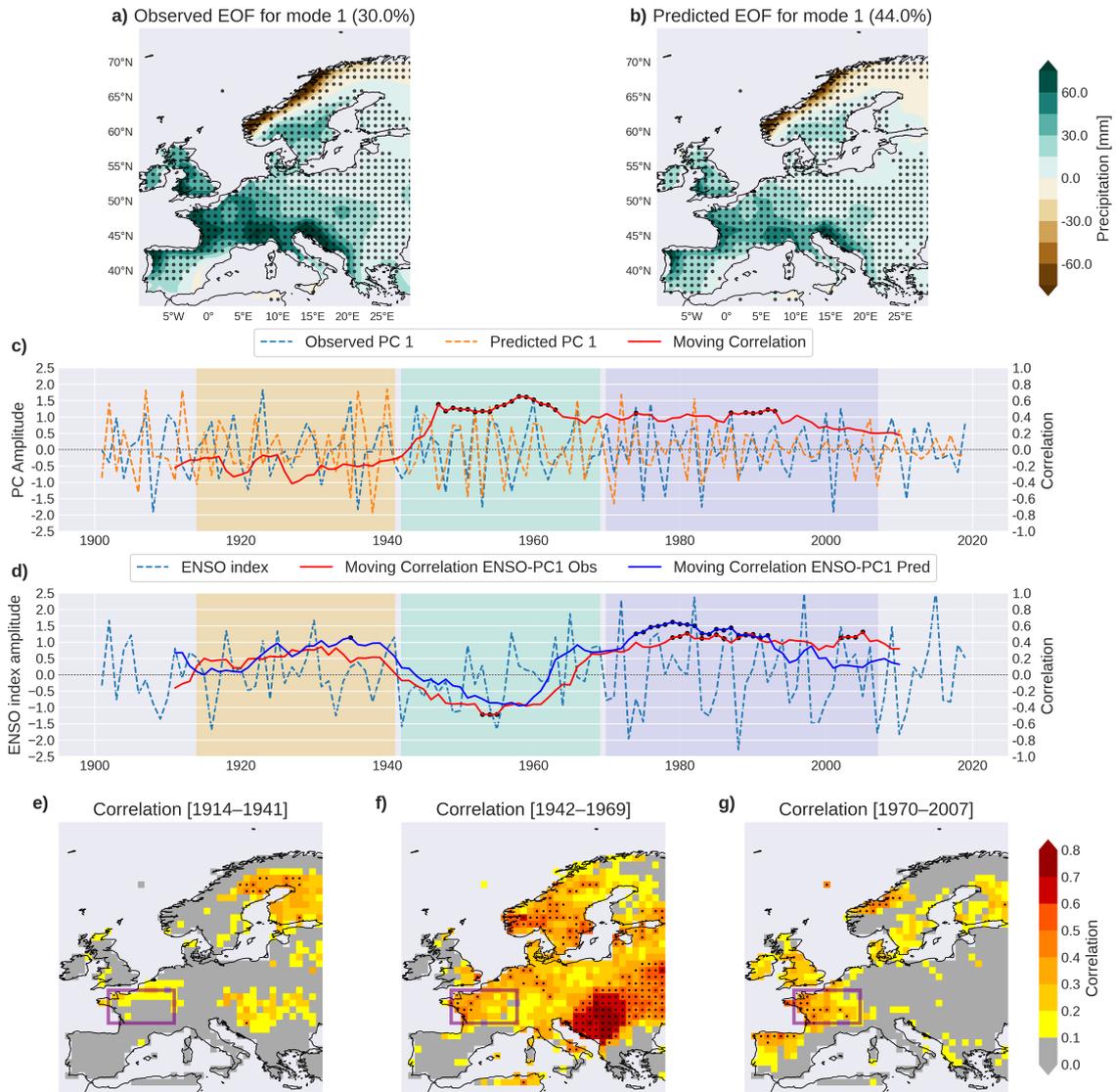
Figure 6. Principal Component Analysis (PCA) of precipitation anomalies over Europe during OND for both observed and predicted datasets. Panels (a) and (b) show the spatial pattern of the first mode of variability for the observed and predicted fields, respectively, with the percentage of explained variance indicated in parentheses. Panel (c) displays the temporal evolution of the principal component (PC) associated with the first mode, representing its high-frequency component after applying a Butterworth filter with a 7-year cut-off, along with the 20-year centered moving window correlation between the observed and predicted PC time series. Panel (d) presents the Niño 3.4 index time series and its 20-year centered moving correlations with both the observed and predicted PCs. Panels (e-g) show the ACC maps between the predicted and observed precipitation high-frequency fields (cut-off of 7 years). Dots in each panel indicate points where the regression/correlation is statistically significant at the 95% confidence level, based on a two-tailed t-test. Three shaded periods in panels (c-d) indicate intervals selected for further analysis following (López-Parages and Rodríguez-Fonseca, 2012).

with ENSO can lead to variations in model reliability. Interestingly, a period of reduced skill appears during 1910-1940, with no significant correlation with El Niño in either the observations or the model. These changes in the reliability of the PCs are also reflected in the skill, as evidenced by the ACC maps shown in panels (e-g) . These results demonstrate that the model not only reproduces the dominant variability of precipitation but also captures the non-stationary nature of the ENSO-European precipitation.

Our results coincide with those of (López-Parages and Rodríguez-Fonseca, 2012), although with minor differences due to changes in data preprocessing. Specifically, while the previous study applied stronger filtering to isolate interannual variability, in the present work we used a Butterworth filter with a cut-off below 7 years, which may retain some lower-frequency variability. Based on the model's ability to reproduce the dominant mode of precipitation variability over Europe (Fig. 6c), we defined three intervals: a low-skill period (1914-1941), a medium-skill period (1970-2007), and a high-kill period (1942-1969). The corresponding composites of model predictions, predictor fields, and attribution maps for positive precipitation events over western-central Europe (purple box in Fig. 7a) are presented in Figure 7. Composites were constructed by selecting events where the respective anomalous precipitation index exceeded ±0.5 standard deviations.

The analysis reveals clear differences between the two periods in the teleconnection mechanisms inferred from the modelled precipitation. During P2, rainfall variability is associated with a weak La Niña and enhanced precipitation over regions such as the Maritime Continent. These anomalies could act as sources of Rossby waves, triggering a circumglobal teleconnection pattern. In contrast, during P3, a strong El Niño signal is present, together with a well-defined Gill-type response and an atmospheric Rossby wave train propagating into the extratropics, impacting precipitation over western-central Europe (Fig. 7f). The attribution maps further clarify the regions contributing to the European precipitation signal. During P2, most of the predictive rainfall signal originates from an extratropical region around 40°N and 160°E, consistent with a weakening of the Aleutian Low and with little evidence of direct influence from the equatorial Pacific. The use of lag-0 relationships could explain contributions from multiple ocean basins, reflecting concurrent atmosphere-ocean interactions influencing rainfall variability. In P3, an additional contribution from the tropical Pacific emerges, consistent with the stronger El Niño signal and the enhanced ENSO influence discussed above. Differences in the background climatology, particularly in the meridional SST gradient and the intensity and position of the jet stream (Fig. 7g-h), help explain the distinct teleconnection patterns: P2 is characterized by a weaker meridional gradient and a southward-displaced jet, producing a weaker Pacific-Europe link that is more closely related to a circumglobal response (Branstator, 2002), whereas P3 exhibits a stronger gradient that enables an enhanced teleconnection. These composites highlight how changes in both tropical Pacific forcing and the extratropical background state contribute to variations in model skill and in the mechanisms linking ENSO to European precipitation, all of which are captured by the simulated rainfall. Composites for negative precipitation events reveal a similar but reversed mechanism (see Fig. S6 in Supplementary Material). We note that the diagnosed relationships represent statistical associations learned by the NN model. While the patterns resemble known modes of variability (e.g., ENSO, NAO), they do not imply causal forcing, which would require dedicated sensitivity experiments to isolate individual contributions.

Additionally, composites based on the observed precipitation index show a dynamical behaviour similar to that of the predicted index (not shown). Regarding the role of the NAO, it is important to note that precipitation is predicted from SSTs, and that
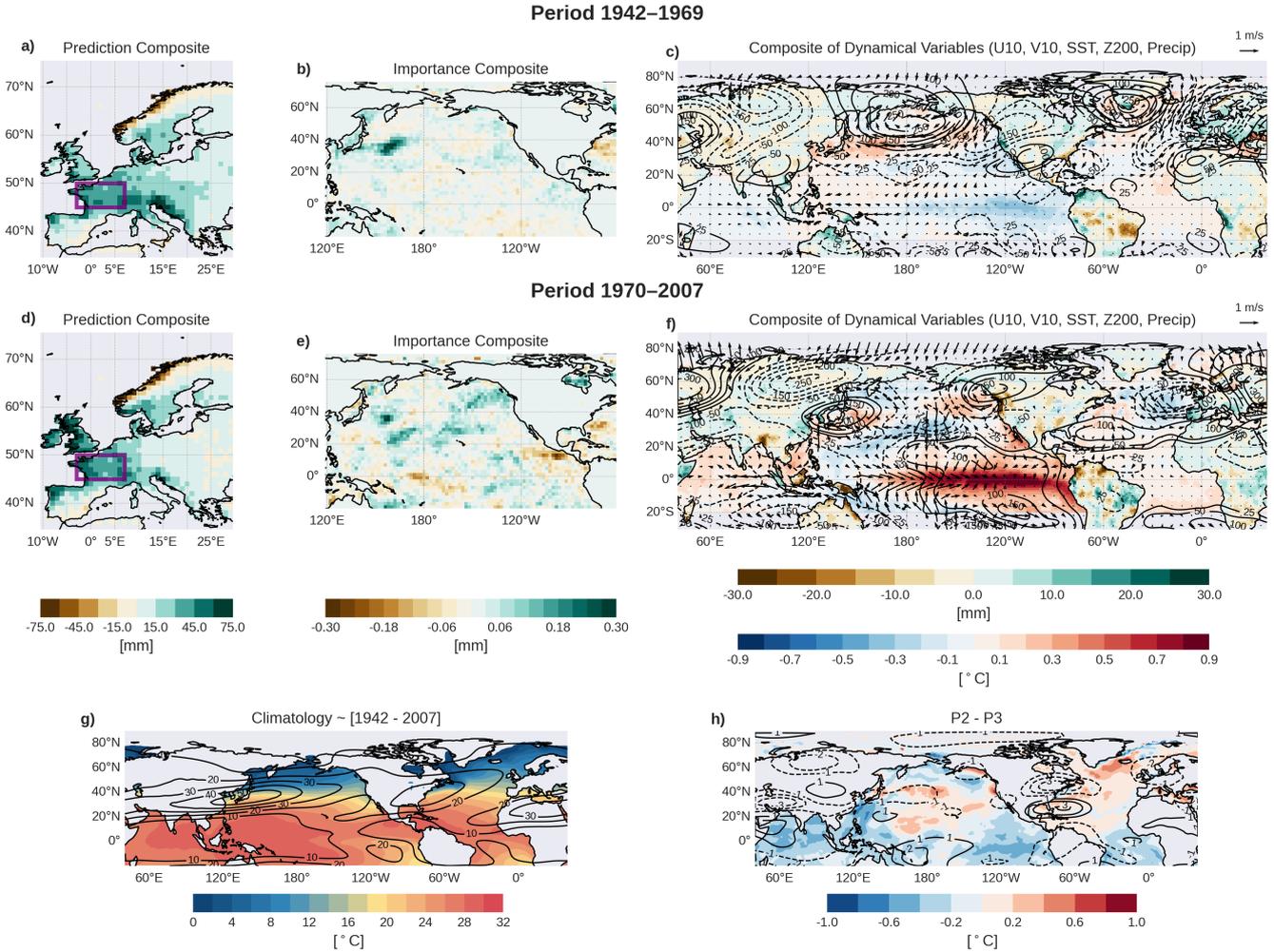
Figure 7. Composites of model predictions, predictor fields, and attribution maps based on positive precipitation events in western central Europe (index defined as the purple rectangle in (a)), with 11 events in period P2 [1942-1969] and 7 events in period P3 [1970-2007]. Panels show: (a, d) precipitation anomalies in Europe during OND for periods P2 and P3, respectively; (b, e) attribution maps over the predictor field corresponding to positive events for periods P2 and P3, respectively. Panels (c, f) display global composites of OND anomalies in sea surface temperature and precipitation (shading), 200 hPa geopotential (contours), and surface winds for positive events for periods P2 and P3, respectively. Attribution maps (b, e) indicate the relative contribution of each grid point in the predictor field to the forecasted value in the target region. The sum of the attribution values within each map equals the predicted anomaly in the corresponding index region (i.e., the sum of values in panel (b) matches the precipitation anomaly within the purple box in panel (a)). Panels (g, h) show the climatology of the SST and U200 for the period [1942-2007] as well as the differences from periods P2 and P3.

20

the ENSO-Europe teleconnection in early winter does not always project onto the NAO. Depending on the period, the ENSO influence can instead manifest as an East Atlantic (EA) pattern (Hou et al., 2023), as observed in P3. There is no clear ENSO forcing on the NAO during P2, when the NAO/EA variability appears to be more internally driven, whereas in P3 it seems to be more strongly forced by SST anomalies. To better identify the drivers of this teleconnection, an analysis of model skill and its attribution as a function of increasing lead time could provide further insight.

5   Conclusions

Seasonal predictions play a crucial role in climate services, offering valuable insights for sectors such as water resource management, energy demand forecasting and agricultural planning, among others. Earth system models make these predictions by taking into account the interactions of the different components of the climate system. Particularly in ocean-atmosphere interactions, the scale of ocean adjustment is slower, which gives predictability to the impacts on the atmosphere. However, seasonal forecasts are subject to different challenges, not only due to the misrepresentation of certain physical processes due to the limited spatio-temporal resolution, but also due to errors in the initial and boundary conditions. An alternative emerges in the statistical approaches trained with observations for performing these simulations. Due to the linear approach of some of the most used statistical models (linear regression, maximum covariance analysis, etc.), deep learning offers an alternative for assessing climate predictability. NN4CAST is specifically designed to overcome some of these limitations. It assesses the sensitivity of a target climate impact variable to variations in drivers operating at seasonal timescales, such as changes in SST anomalies across different regions, and identify the regions from which the model extracts the information.

The present paper demonstrate, by analizing two state of the art teleconnections, the advantages of this modelling framework, not just for producing retrospective hindcasts but also to analyze the sources of skill and identify the regions that gives predictability to the system. NN4CAST primary applications include:

· Providing a versatile and user-friendly tool for ease of implementation and application.

· Modelling relationships between fields, including non-linear components, with performance evaluated using various skill and error metrics.

· Assessing potential predictors and exploring the impact of target variables on specific drivers to enhance predictability.

· Identifying windows of opportunity where relationships between fields exhibit stronger predictability.

· Facilitating the analysis of changes in predictability by examining the model attributions in the predictions.

Two case studies, based on well-known climate teleconnections, have been selected to analyse the potential applications of the proposed tool and pose new hypothesis that could be inferred thanks to this approach. First, the potential predictability of tropical north Atlantic SSTs is assessed training the model with DJF tropical Pacific SSTs, due to the fact of the known that there is a robust teleconnection between ENSO in boreal winter (DJF) and the following anomalous SSTs in the TNA. Then, the sources of predictability are analysed together with the atmospheric teleconnections that can be inferred from the predicted field.

21

Second, OND seasonal European rainfall simulated by the model training with the contemporaneous SSTs are used to infer hypothesis about the sources of predictability and teleconnection mechanisms, due to the fact that this teleconnection has been found to be non-stationary and change at decadal timescales. In this context, the model enables the identification of periods of enhanced predictability and provides insights into the potential drivers of such variability. In our first case study, examining the ENSO-TNA SST teleconnection, the model successfully reproduces the canonical relationship with high skill and, through attribution analysis, identifies the central Pacific as a key region of influence that traditional regression methods fail to detect. In the second case study, which assesses the impact of Pacific SSTs on European precipitation, the model accurately captures the non-stationarity behaviour of the teleconnection, achieving the highest skill during the period of negative ENSO-European precipitation correlation and reproducing the dynamical patterns characteristic of that phase.

The case studies presented demonstrate that NN4CAST enables model assessment through performance metrics and visualization tools, which can be used to identify windows of opportunity (WoO) for seasonal prediction. For example, in the European precipitation case study, period P2 (1942-1969) represent a WoO, where predictive skill is high for predicting precipitation. This illustrates how NN4CAST can guide the identification of time intervals in which forecasts are most reliable, supporting more targeted and effective seasonal prediction strategies. Model skill metrics and the analysis of attributions in the predictive field, outcomes of NN4CAST, allow users to better understand climate teleconnections and its underlying physical mechanisms.

The NN4CAST framework further facilitates the implementation of pseudo-sensitivity experiments, whereby users can, for instance, select different SST regions as predictors and evaluate their individual and combined contributions by applying them both jointly and separately. This capability, together with the possibility to vary predictands, and employ different datasets, enhances the framework utility for systematically assessing model sensitivities and robustness. Additionally, this framework is designed to allow the integration of more complex deep learning architectures currently used in meteorological modelling, such as transformers. This capability enables a direct evaluation of the strengths and limitations of each architecture within the modelling framework. Furthermore, plans are underway to extend the NN4CAST framework to additional applications relevant to the scientific community, including integration with tools such as ESMValTool (Righi et al., 2020), thereby enhancing its utility for comprehensive climate model evaluation and diagnostics.

References

460 Alexander, M. A., Bladé, I., Newman, M., Lanzante, J. R., Lau, N.-C., and Scott, J. D.: The atmospheric bridge: The influence of ENSO teleconnections on air–sea interaction over the global oceans, Journal of climate, 15, 2205–2231, 2002.

Alvarez-Ramirez, J., Rodriguez, E., and Echeverría, J. C.: Detrending fluctuation analysis based on moving average filtering, Physica A: statistical mechanics and its applications, 354, 199–219, 2005.

Bachèlery, M.-L., Brajard, J., Patacchiola, M., Illig, S., and Keenlyside, N.: Predicting Atlantic and Benguela Niño events with deep learning,
465 Science Advances, 11, eads5185, 2025.

Bergstra, J. and Bengio, Y.: Random search for hyper-parameter optimization., Journal of machine learning research, 13, 2012.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q.: Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast, arXiv preprint arXiv:2211.02556, 2022.

Branstator, G.: Circumglobal teleconnections, the jet stream waveguide, and the North Atlantic Oscillation, Journal of Climate, 15, 1893–1910,
470 2002.

Calvo-Miguélez, E., Rodríguez-Fonseca, B., Galván-Fraile, V., and Gómara, I.: Predicting Chlorophyll-A in the Mauritanian-Senegalese Coastal Upwelling from Tropical Sea Surface Temperature, 2025.

Chang, P., Ji, L., and Saravanan, R.: A hybrid coupled model study of tropical Atlantic variability, Journal of climate, 14, 361–390, 2001.

Cuomo, S., Di Cola, V. S., Giampaolo, F., Rozza, G., Raissi, M., and Piccialli, F.: Scientific machine learning through physics–informed neural
475 networks: Where we are and what's next, Journal of Scientific Computing, 92, 88, 2022.

Czaja, A., Van der Vaart, P., and Marshall, J.: A diagnostic study of the role of remote forcing in tropical Atlantic variability, Journal of Climate, 15, 3280–3290, 2002.

Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P., and Rodrigues, L. R.: Seasonal climate predictability and forecasting: status and prospects, Wiley Interdisciplinary Reviews: Climate Change, 4, 245–268, 2013.

480 Enfield, D. B. and Mayer, D. A.: Tropical Atlantic sea surface temperature variability and its relation to El Niño-Southern Oscillation, Journal of Geophysical Research: Oceans, 102, 929–945, 1997.

Galván Fraile, V., Martín-Rey, M., Rodríguez-Fonseca, B., Polo, I., and Navarro-García, M.: NN4CASTv1.0.20, Zenodo [code], https://doi.org/10.5281/zenodo.14011998, 2024.

Galván Fraile, V., Martín-Rey, M., Rodríguez-Fonseca, B., Polo, I., and Navarro-García, M.: NN4CAST_manual, Zenodo [simulation outputs
485 and script], https://doi.org/10.5281/zenodo.17287629, 2025.

García-Serrano, J., Cassou, C., Douville, H., Giannini, A., and Doblas-Reyes, F. J.: Revisiting the ENSO teleconnection to the tropical North Atlantic, Journal of Climate, 30, 6945–6957, 2017.

Géron, A.: Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow, " O'Reilly Media, Inc.", 2022.

Gill, A. E.: Some simple solutions for heat-induced tropical circulation, Quarterly Journal of the Royal Meteorological Society, 106, 447–462,
490 1980.

Gleckler, P. J., Taylor, K. E., and Doutriaux, C.: Performance metrics for climate models, Journal of Geophysical Research, 113, 2008.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y.: Deep learning, vol. 1, MIT press Cambridge, 2016.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D.: A survey of methods for explaining black box models, ACM computing surveys (CSUR), 51, 1–42, 2018.

495 Ham, Y.-G., Kim, J.-H., and Luo, J.-J.: Deep learning for multi-year ENSO forecasts, Nature, 573, 568–572, 2019.

Hargreaves, J. C.: Skill and uncertainty in climate models, Wiley Interdisciplinary Reviews: Climate Change, 1, 556–564, 2010.

Harris, I., Osborn, T. J., Jones, P., and Lister, D.: Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset, Scientific data, 7, 109, 2020.

He, K., Zhang, X., Ren, S., and Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE international conference on computer vision, pp. 1026–1034, 2015.

Hinton, G. E. and Salakhutdinov, R. R.: Reducing the dimensionality of data with neural networks, science, 313, 504–507, 2006.

Horel, J. D. and Wallace, J. M.: Planetary-scale atmospheric phenomena associated with the Southern Oscillation, Monthly Weather Review, 109, 813–829, 1981.

Hou, J., Fang, Z., and Geng, X.: Recent Strengthening of the ENSO Influence on the Early Winter East Atlantic Pattern, Atmosphere, 14, 1809, 2023.

Ibebuchi, C. C. and Richman, M. B.: Deep learning with autoencoders and LSTM for ENSO forecasting, Climate Dynamics, 62, 5683–5697, 2024.

Ineson, S. and Scaife, A.: The role of the stratosphere in the European climate response to El Niño, Nature Geoscience, 2, 32–36, 2009.

Johnson, S. J., Stockdale, T. N., Ferranti, L., Balmaseda, M. A., Molteni, F., Magnusson, L., Tietsche, S., Decremer, D., Weisheimer, A., Balsamo, G., et al.: SEAS5: the new ECMWF seasonal forecast system, Geoscientific Model Development, 12, 1087–1117, 2019.

Kent, C., Scaife, A. A., Dunstone, N. J., Smith, D., Hardiman, S. C., Dunstan, T., and Watt-Meyer, O.: Skilful global seasonal predictions from a machine learning weather model trained on reanalysis data, npj Climate and Atmospheric Science, 8, 314, 2025.

Kirtman, B. and Pirani, A.: The state of the art of seasonal prediction: Outcomes and recommendations from the First World Climate Research Program Workshop on Seasonal Prediction, Bulletin of the American Meteorological Society, 90, 455–458, 2009.

Klaise, J., Van Looveren, A., Vacanti, G., and Coca, A.: Alibi explain: Algorithms for explaining machine learning models, Journal of Machine Learning Research, 22, 1–7, 2021.

Klein, S. A., Soden, B. J., and Lau, N.-C.: Remote sea surface temperature variations during ENSO: Evidence for a tropical atmospheric bridge, Journal of climate, 12, 917–932, 1999.

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., Klöwer, M., Lottes, J., Rasp, S., Düben, P., et al.: Neural general circulation models for weather and climate, Nature, 632, 1060–1066, 2024.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., et al.: GraphCast: Learning skillful medium-range global weather forecasting, arXiv preprint arXiv:2212.12794, 2022.

LeCun, Y., Bengio, Y., and Hinton, G.: Deep learning, nature, 521, 436–444, 2015.

Lee, S.-K., Enfield, D. B., and Wang, C.: Why do some El Niños have no impact on tropical North Atlantic SST?, Geophysical Research Letters, 35, 2008.

López-Parages, J. and Rodríguez-Fonseca, B.: Multidecadal modulation of El Niño influence on the Euro-Mediterranean rainfall, Geophysical Research Letters, 39, 2012.

López-Parages, J., Rodríguez-Fonseca, B., and Terray, L.: A mechanism for the multidecadal modulation of ENSO teleconnection with Europe, Climate Dynamics, 45, 867–880, 2015.

Mamalakis, A., Barnes, E. A., and Ebert-Uphoff, I.: Investigating the fidelity of explainable artificial intelligence methods for applications of convolutional neural networks in geoscience, Artificial Intelligence for the Earth Systems, 1, e220 012, 2022.

Michaelsen, J.: Cross-validation in statistical climate forecast models, Journal of Applied Meteorology and Climatology, 26, 1589–1600, 1987.

Pan, B., Anderson, G. J., Goncalves, A., Lucas, D. D., Bonfils, C. J., and Lee, J.: Improving seasonal forecast using probabilistic deep learning, Journal of Advances in Modeling Earth Systems, 14, e2021MS002 766, 2022.

535 Poli, P., Hersbach, H., Dee, D. P., Berrisford, P., Simmons, A. J., Vitart, F., Laloyaux, P., Tan, D. G., Peubey, C., Thépaut, J.-N., et al.: ERA-20C: An atmospheric reanalysis of the twentieth century, Journal of Climate, 29, 4083–4097, 2016.

Raffalovich, L. E.: Detrending time series: A cautionary note, Sociological Methods & Research, 22, 492–519, 1994.

Rawat, W. and Wang, Z.: Deep convolutional neural networks for image classification: A comprehensive review, Neural computation, 29, 2352–2449, 2017.

540 Rayner, N., Parker, D. E., Horton, E., Folland, C. K., Alexander, L. V., Rowell, D., Kent, E. C., and Kaplan, A.: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century, Journal of Geophysical Research, 108, 2003.

Rew, R., Hartnett, E., Caron, J., et al.: NetCDF-4: Software implementing an enhanced data model for the geosciences, in: 22nd International Conference on Interactive Information Processing Systems for Meteorology, Oceanograph, and Hydrology, vol. 6, 2006.

Richter, I. and Tokinaga, H.: An overview of the performance of CMIP6 models in the tropical Atlantic: mean state, variability, and remote

545 impacts, Climate Dynamics, 55, 2579–2601, 2020.

Rieger, N., Corral, Á., Olmedo, E., and Turiel, A.: Lagged teleconnections of climate variables identified via complex rotated maximum covariance analysis, Journal of Climate, 34, 9861–9878, 2021.

Righi, M., Andela, B., Eyring, V., Lauer, A., Predoi, V., Schlund, M., Vegas-Regidor, J., Bock, L., Brötz, B., de Mora, L., et al.: Earth System model evaluation tool (ESMValTool) v2. 0–technical overview, Geoscientific Model Development, 13, 1179–1199, 2020.

550 Roberts, C., Vitart, F., and Balmaseda, M.: Hemispheric impact of North Atlantic SSTs in subseasonal forecasts, Geophysical Research Letters, 48, e2020GL0911 446, 2021.

Rodríguez-Fonseca, B., Suárez-Moreno, R., Ayarzagüena, B., López-Parages, J., Gómara, I., Villamayor, J., Mohino, E., Losada, T., and Castaño-Tierno, A.: A review of ENSO influence on the North Atlantic. A non-stationary signal, Atmosphere, 7, 87, 2016.

Rodwell, M. J., Rowell, D. P., and Folland, C. K.: Oceanic forcing of the wintertime North Atlantic Oscillation and European climate, Nature,

555 398, 320–323, 1999.

Rogers, J. C.: North Atlantic storm track variability and its association to the North Atlantic Oscillation and climate variability of northern Europe, Journal of Climate, 10, 1635–1647, 1997.

Ronneberger, O., Fischer, P., and Brox, T.: U-net: Convolutional networks for biomedical image segmentation, in: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part

560 III 18, pp. 234–241, Springer, 2015.

Russell, S. J. and Norvig, P.: Artificial intelligence: a modern approach, Pearson, 2016.

Samuel, A. L.: Some studies in machine learning using the game of checkers, IBM Journal of research and development, 3, 210–229, 1959.

Schulzweida, U., Kornblueh, L., and Quast, R.: CDO user guide, 2019.

Shin, N.-Y., Ham, Y.-G., Kim, J.-H., Cho, M., and Kug, J.-S.: Application of deep learning to understanding ENSO dynamics, Artificial

565 Intelligence for the Earth Systems, 1, e210 011, 2022.

Shuila, J. and Kinter III, J.: Predictability of seasonal climate variations: A pedagogical review, 2006.

Suárez-Moreno, R. and Rodríguez-Fonseca, B.: S^{4}CAST v2.0: sea surface temperature based statistical seasonal forecast model, Geoscientific model development, 8, 3639–3658, 2015.

Sundararajan, M., Taly, A., and Yan, Q.: Axiomatic attribution for deep networks, in: International conference on machine learning, pp.

570 3319–3328, PMLR, 2017.

Trigo, R. M., Osborn, T. J., and Corte-Real, J. M.: The North Atlantic Oscillation influence on Europe: climate impacts and associated physical mechanisms, Climate research, 20, 9–17, 2002.

Wade, M., Rodríguez-Fonseca, B., Martín-Rey, M., Lazar, A., López-Parages, J., and Gaye, A. T.: Interdecadal changes in SST variability drivers in the Senegalese-upwelling: the impact of ENSO, Climate Dynamics, 60, 667–685, 2023.

575 Watt-Meyer, O., Henn, B., McGibbon, J., Clark, S. K., Kwa, A., Perkins, W. A., Wu, E., Harris, L., and Bretherton, C. S.: ACE2: Accurately learning subseasonal to decadal atmospheric variability and forced responses, arXiv preprint arXiv:2411.11268, 2024.

Weisheimer, A., Schaller, N., O'Reilly, C., MacLeod, D. A., and Palmer, T.: Atmospheric seasonal forecasts of the twentieth century: multi-decadal variability in predictive skill of the winter North Atlantic Oscillation (NAO) and their potential value for extreme event attribution, Quarterly Journal of the Royal Meteorological Society, 143, 917–926, 2017.

580 Wilks, D. S.: Improved statistical seasonal forecasts using extended training data, International Journal of Climatology: A Journal of the Royal Meteorological Society, 28, 1589–1598, 2008.

Wilks, D. S.: Statistical methods in the atmospheric sciences, vol. 100, Academic press, 2011.

Wilks, D. S.: Comparison of probabilistic statistical forecast and trend adjustment methods for North American seasonal temperatures, Journal of Applied Meteorology and Climatology, 53, 935–949, 2014.

585 Wu, R., Lin, M., and Sun, H.: Impacts of different types of El Niño and La Niña on northern tropical Atlantic sea surface temperature, Climate Dynamics, 54, 4147–4167, 2020.

Supplementary Material

Hyperparameter Dictionary and Preprocessing Code

NN4CAST stores model configurations in a Python dictionary, which can be saved as a YAML file to facilitate reproducibility. The following examples demonstrate the creation of this dictionary, saving it, and preprocessing the data.

```
     from tensorflow.keras import activations
     import numpy as np
     from nn4cast.predefined_classes import Dictionary_saver

5    hyperparameters =
         'path_x' = "/path/to/your/data/HadISST1_sst_1870-2019.nc",
         'path_y' = "/path/to/your/data/HadISST1_sst_1870-2019.nc",

         'time_limits' = [1900,2019],
10       'jump_year' = 0,

         'lat_lims_x' = [30, -30],
         'lon_lims_x' = [120, 290],
         'lat_lims_y' = [40, -10],
15       'lon_lims_y' = [-80, +20],

         'name_x' = 'sst',
         'name_y' = 'sst',

20       'months_x' = [1, 2, 3],
         'months_skip_x' = ['None'],
         'months_y' = [4, 5, 6],
         'months_skip_y' = ['None'],

25       'mean_seasonal_method_x' = True,
         'mean_seasonal_method_y' = True,

         'regrid_degree_x' = 2,
         'regrid_degree_y' = 2,
30
         'scale_x' = 1,
         'scale_y' = 1,

         'detrend_x' = True,
35       'detrend_y' = True,
         'detrend_x_window' = 30,
         'detrend_y_window' = 30,
         # Neural network hyperparameters (default ones).
         'layer_sizes' = [1024,256,64,256,1024],
40       'activations' = [activations.elu,activations.elu,activations.elu,
           activations.elu, activations.elu]
         'dropout_rates' = [0.1],
```

```
         'kernel_regularizer' = 'l2',
         'learning_rate' = 0.0001,
         'epochs' = 2500,
         'num_conv_layers' = 0,
57       'use_batch_norm' = True,
         'use_initializer' = True,
         'use_dropout' = True,
         'use_init_skip_connections' = False,
         'use_inter_skip_connections' = False,
62
         'units_x' = ['°C'],
         'units_y' = ['°C'],
         'region_predictor' = 'Tropical Pacific',
         'p_value' = 0.1,
67       'outputs_path' = "/path/to/the/directory/Outputs_ND_sst_SO/"

       Dictionary_saver(hyperparameters)
```

Listing S1. Python code example demonstrating how to create an instance of the parameter and hyperparameter dictionary, and how to save it using the `Dictionary-saver` function.

5   The preprocessing of predictor and predictand data is carried out with the `Preprocess` function, which executes the operations defined in the hyperparameter dictionary, including regridding, computing seasonal anomalies, and detrending:

```
       from nn4cast.predefined_classes import Preprocess

       dictionary_preprocess = Preprocess(dictionary_hyperparams=hyperparameters)
```

Listing S2. Python code example demonstrating the use of the `Preprocess` function to manipulate predictor and predictand data, and to store the processed outputs in a dictionary format.

Detrending is implemented with a backward moving average (BMA) algorithm, where the mean over a sliding window of previous years is subtracted from each subsequent value. The window length is specified in the hyperparameters as `detrend_x_window` and `detrend_y_window`. This method ensures that only past information is used and avoids
10 introducing future knowledge into the preprocessing phase.

Model Cross-Validation and Explainable AI

NN4CAST implements a cross-validation procedure using the `Model_build_and_test()` function. The number of folds is defined by `n_cv_folds`, and a fraction of each training fold (10%) is reserved for early stopping. This procedure ensures robust estimates of model performance and can be configured for leave-one-out cross-validation if desired. Attributions for a
15 target region are computed using the Integrated Gradients methodology from the Alibi library (Klaise et al., 2021), leveraging

the cross-validation results to cover the entire hindcast period.To compute these attributions for a specific target region of the predictand field, first set the instance to `True` to enable the computation. Then, define the region as a list of latitude and longitude ranges, using the parameter (setting region_importances=[[latitude_min, latitude_max], [longitude_min, longitude_max]]).

```python
from nn4cast.predefined_classes import Model_build_and_test, Results_plotter

outputs_cross_validation = Model_build_and_test(dictionary_hyperparams =
    hyperparameters, dictionary_preprocess=dictionary_preprocess,
    cross_validation=True, n_cv_folds=120, plot_differences=False, importances=True,
    region_importances=[[5,25],[-55,-15]])

Results_plotter(hyperparameters, dictionary_preprocess, rang_x=1.5, rang_y=1,
    predictions=outputs_cross_validation['predictions'],
    observations=outputs_cross_validation['observations'],
    years_to_plot=[2015], plot_with_contours=True,
    importances=outputs_cross_validation['importances'],
    region_importances=outputs_cross_validation['region_attributed'])
```

Listing S3. Python code example demonstrating the use of the `Model-build-and-test` function to construct the model using selected hyperparameters, followed by training, validation, and testing through a user-defined cross-validation approach. The listing also includes the use of the `Results-plotter` function to visualize the model outputs and their attributions for the target region over the predictor field..

Hyperparameter Optimization

20 Hyperparameter tuning is carried out with the `Model_searcher()` function, which explores a user-defined search space using a random search methodology (Bergstra and Bengio, 2012). The best combination is selected based on cross-validation performance.

```
from nn4cast.predefined_classes import Model_searcher

params_selection =
    'pos_number_layers' = 5, # set the maximum value of fully connected layers (int).
    'pos_layer_sizes' = [16, 64, 256],# set the possible layer sizes (list).
    'pos_activations' = ["elu", "linear"],# set the possible activation functions
        (possibilities are all the ones availabe: tf.keras.layers.activations()) (list).
    'pos_dropout' = [0.0, 0.01],# set the possible dropout probabilities (list).
    'pos_kernel_regularizer' = ["l1_l2"],# set the possible kernel regularizer
        (possibilities are: l1_l2, l1, l2, None) (list).
    'search_skip_connections' = False,# set if searching for skip connections, either
        intermediate or end_to_end connections (bool).
    'pos_conv_layers' = 0,# set the maximum number of convolutional layers, the
        predictor field (X) must be 2D (int).
    'pos_learning_rate' = [1e-4,1e-3],# set the possible learning rates (list).

outputs_bm_cross_validation = Model_searcher(dictionary_hyperparams=hyperparameters,
    dictionary_preprocess=dictionary_preprocess,
    dictionary_possibilities=params_selection, max_trials=10, n_cv_folds=8)
```

Listing S4. Python code example demonstrating the use of the `Model-searcher` function to identify the optimal set of hyperparameters within a predefined search space. The hyperparameter ranges are based on commonly reported values in the specialized literature and explored using the random search method (Bergstra and Bengio, 2012). The best combination obtained is then evaluated following the selected cross-validation scheme.

Application 1: Modelling Pacific-North Tropical Atlantic SST Connection

The `Results-plotter` function enables visualization of the model predictions for specific cases, allowing for direct comparison with the observed data. Two representative years—1986 and 1998—have been selected for this purpose and are illustrated in Fig. S1 and Fig. S2, respectively.
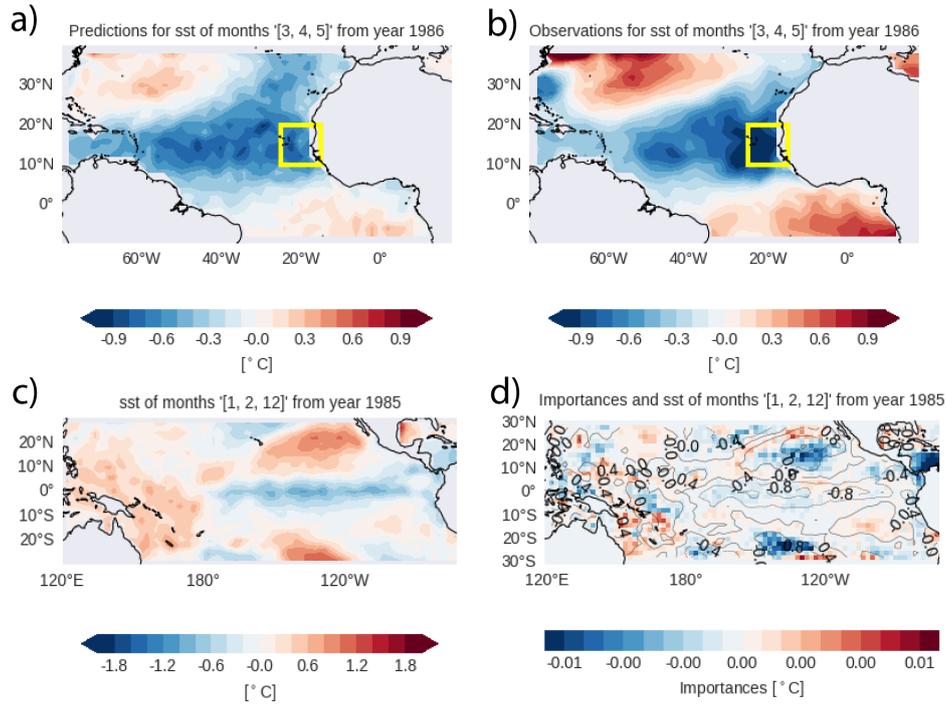
Figure S1. Comparison of model predictions for a specific sample year. Panels show: (a) predicted MAM Atlantic SST anomalies (°C); (b) observed MAM Atlantic SST anomalies (°C) for the year 1986; (c) Pacific SST during DJF (°C) for the year 1985 used as the predictor field; and (d) Pacific SST DJF predictor field for the 1985 sample, with contours (°C) representing SST values and shaded areas (°C) indicating the model attributions for the selected predictand region (yellow box in panels a and b). The yellow box marks the target region where model attributions are computed.

The first case presented (Fig. S1) corresponds to a central La Niña event, characterized by negative SST anomalies in the central equatorial Pacific, along with positive anomalies over the Warm Pool and subtropical regions of the Pacific Ocean. In this scenario, the model predicted a cooling of the WTNA and the SMSCU region, as well as a warming over the eastern equatorial Atlantic and the Gulf Stream region. To gain insight into how the model made these predictions, we focused on the SMSCU region (highlighted by a yellow box in Fig. S1a) and Fig. S1b) and computed the Integrated Gradients attributions for the 1986 prediction. These attributions are shown in Fig. S1d). The importance values are expressed in units of the predicted variable (°C), which in this case matches the units of the predictor (SST anomalies). The attribution map reveals the individual contribution of each SST predictor to the prediction over the region marked by the yellow box. Notably, both the Caribbean region and the warming zones in the central tropical Pacific greatly contributed to the prediction of negative SST anomalies in the Mauritanian-Senegalese Coastal Upwelling region.
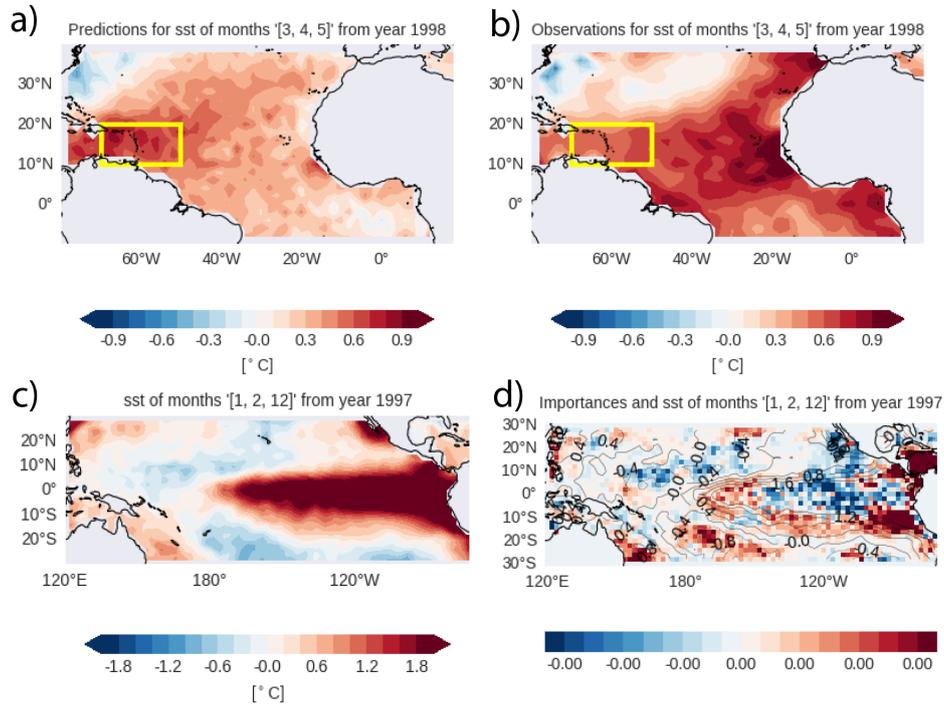
5

Figure S2. Comparison of model predictions for a specific sample year. Panels show: (a) predicted MAM Atlantic SST anomalies (°C); (b) observed MAM Atlantic SST anomalies (°C) for the year 1998; (c) Pacific SST during DJF (°C) for the year 1997 used as the predictor field; and (d) Pacific SST DJF predictor field for the 1997 sample, with contours (°C) representing SST values and shaded areas (°C) indicating the model attributions for the selected predictand region (yellow box in panels a and b). The yellow box marks the target region where model attributions are computed.

Figure S2 illustrates the second case, corresponding to a canonical El Niño event in 1998, associated with widespread warming in the tropical Atlantic, including the WTNA region. In this case, the model accurately reproduced the observed warming signal. For the attribution analysis, we selected a classic WTNA index region, delineated by the yellow box in Fig. S2a) and Fig. S2b). The attribution map for this case is somewhat noisier compared to the 1986 scenario. Nevertheless, some consistent patterns emerge: there are positive contributions from the Caribbean, the central Pacific, and a latitudinal band south of 10°S. In contrast, negative contributions are found mainly in the Niño 3 region and certain areas north of 10°N, which act to counterbalance the warming signal.
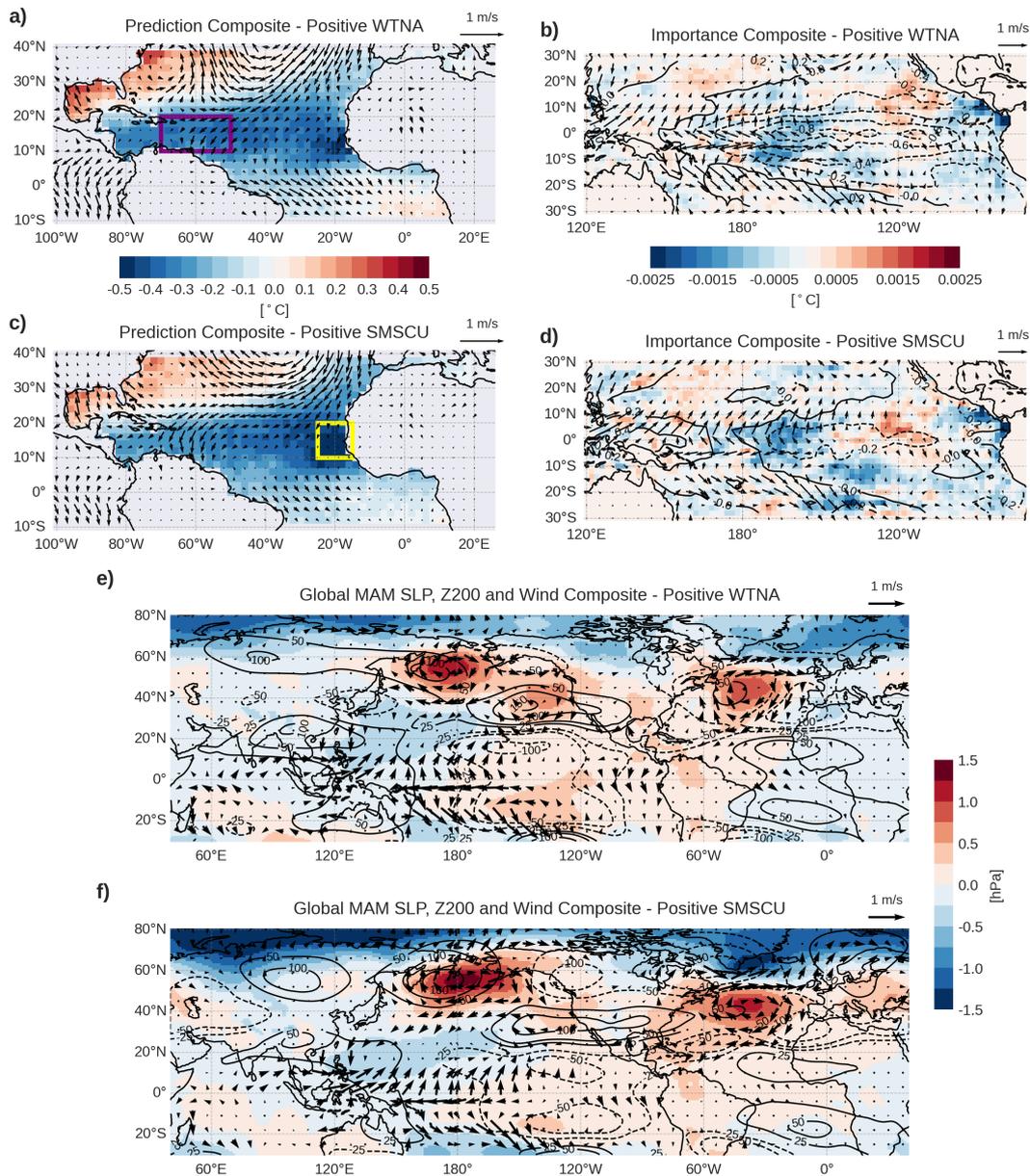
Figure S3. Composites of model anomalous SST predictions, predictor fields, and attribution maps for negative predicted WTNA and SMSCU, based on 28 and 29 events, respectively. Panels (a) and (c) show the predicted mean SST anomalies in the Atlantic during MAM together with surface wind anomalies indicated by arrows. Panels (b) and (d) show the attribution maps over the predictor fields with SST in contours and surface winds in arrows. Panels (e) and (f) display global composites of MAM anomalies in sea level pressure (shading), 200 hPa geopotential (contours), and surface winds for positive WTNA and SMSCU events. Attribution maps indicate the relative contribution of each grid point in the predictor field to the forecasted value in the target region, with the sum of the values within each map matching the predicted anomaly in the corresponding index region (i.e., the sum of values in panel (c) matches the WTNA anomaly within the purple box in panel (a)).

Figure S3 represents the same as Figure 4 but for the negative events. Composites based on the Niño 3.4 index are also presented in Figure S4 A traditional approach of teleconnection analysis is the regression of the indices over the SST fields. This is illustrated for comparison in Figure S5.
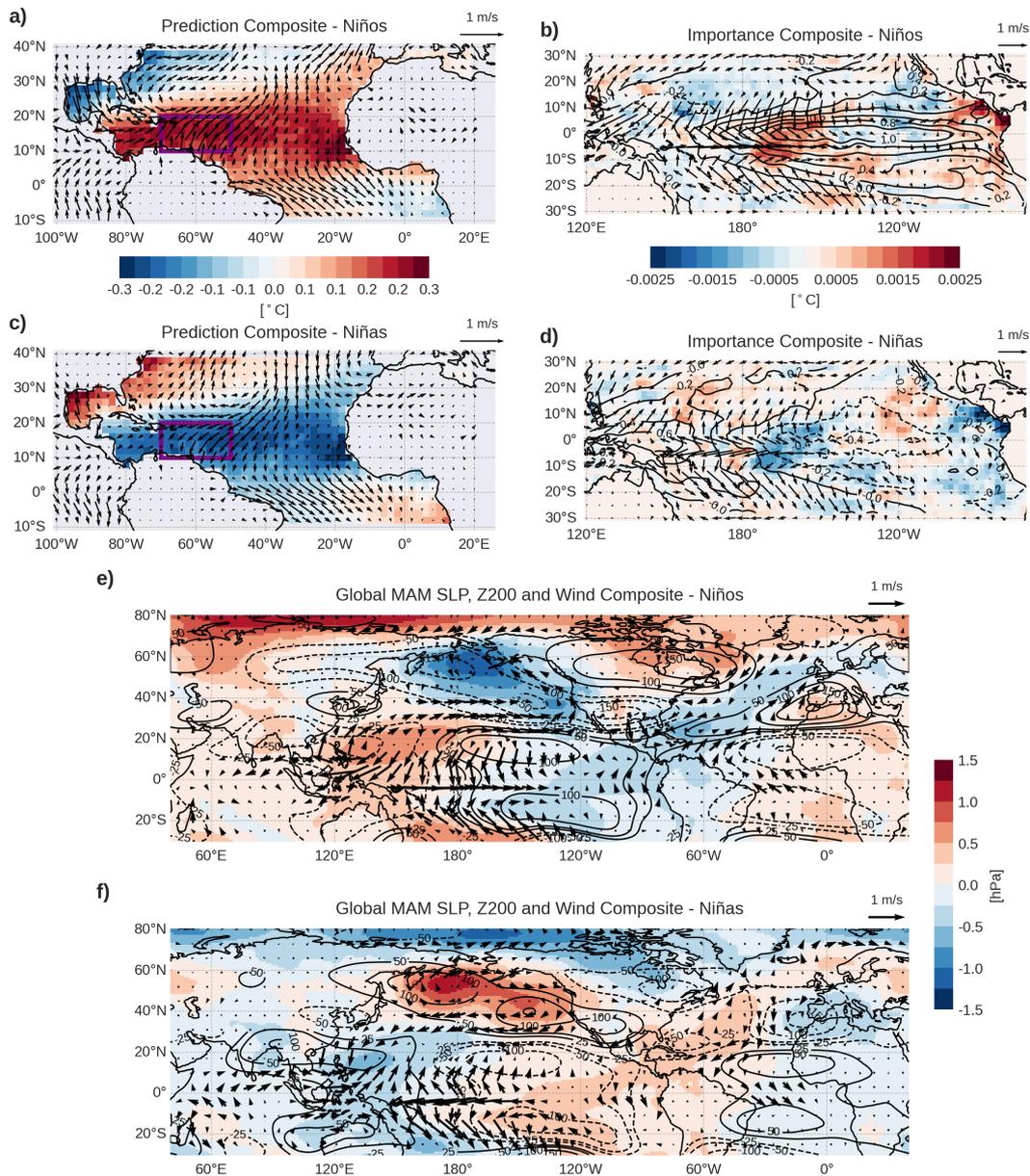
Figure S4. Composites of model anomalous SST predictions, predictor fields, and attribution maps for Niño 3.4 positive (Niño) and negative (Niña), based on 35 and 34 events respectively. Panels (a) and (c) show the predicted mean SST anomalies in the Atlantic during MAM together with surface wind anomalies indicated by arrows. Panels (b) and (d) show the attribution maps over the predictor fields with SST in contours and surface winds in arrows. Panels (e) and (f) display global composites of MAM anomalies in sea level pressure (shading), 200 hPa geopotential (contours), and surface winds for Niños and Niñas. Attribution maps indicate the relative contribution of each grid point in the predictor field to the forecasted value in the target region, with the sum of the values within each map matching the predicted anomaly in the corresponding index region (i.e., the sum of values in panel c matches the WTNA anomaly within the purple box in panel (a)).
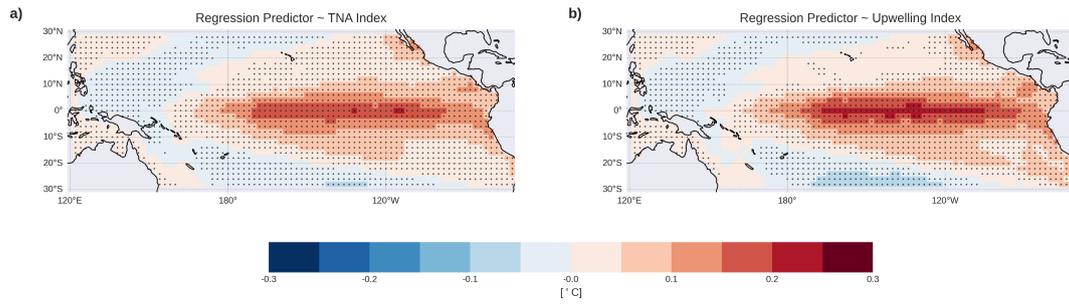
Figure S5. Regressions of WNTA and SMSCU indices in MAM over the DJF SST anomalies. Statistically significant results, based on a two-tailed t-test with a significance level defined in the hyperparameters (95%), are indicated by the dotted regions.

Application 2: Modelling North and Tropical Pacific - European Precipitation Teleconnection

Figure S6 shows the same information as Figure 7, but for the composite of predicted negative precipitation anomalies.
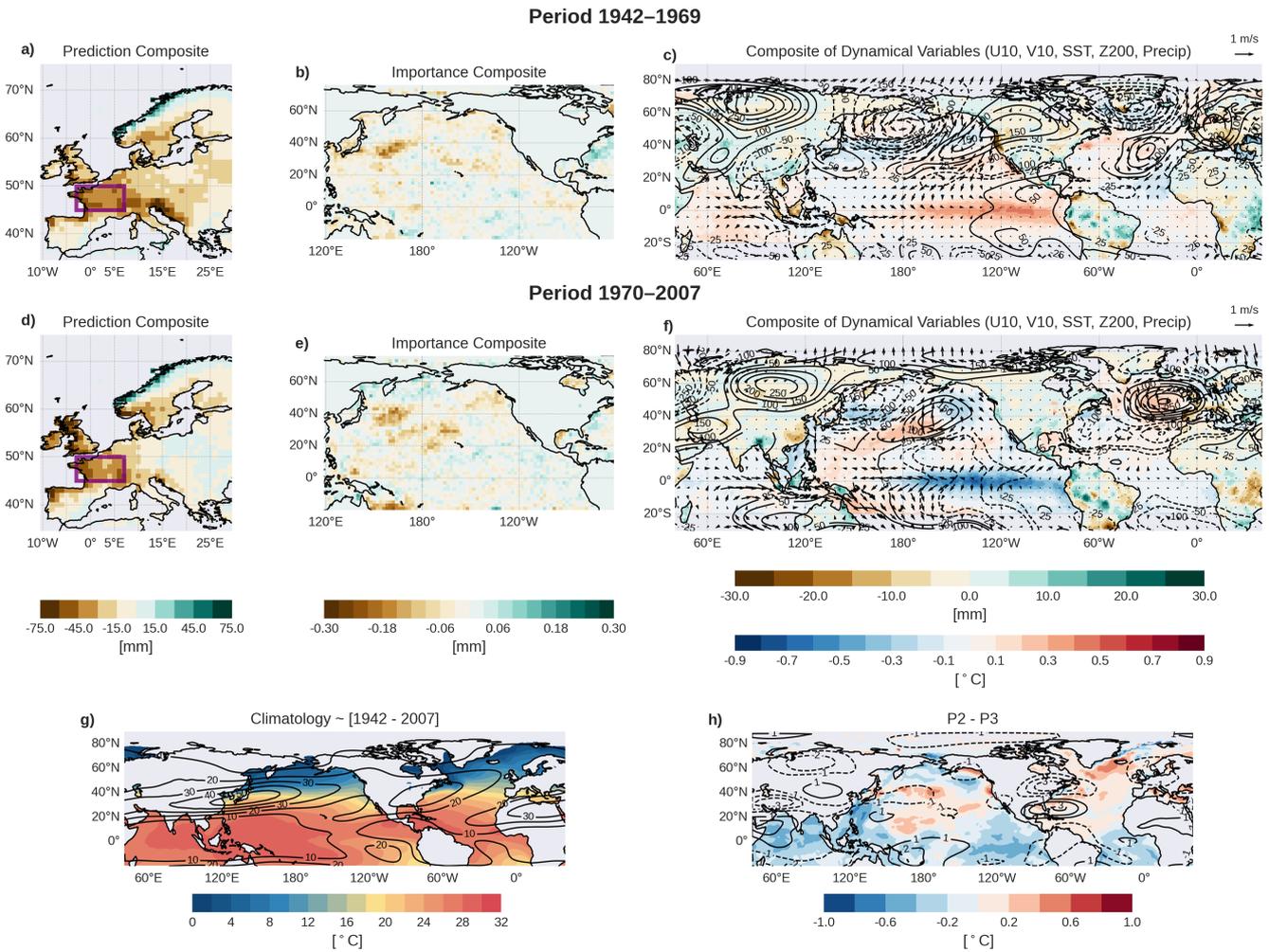
Figure S6. Composites of model predictions, predictor fields, and attribution maps based on negative precipitation events in western central Europe (index defined as the purple rectangle in (a)), with 11 events in period P2 [1942-1969] and 11 events in period P3 [1970-2007]. Panels show: (a, d) precipitation anomalies in Europe during OND for periods P2 and P3, respectively; (b, e) attribution maps over the predictor field corresponding to negative events for periods P2 and P3, respectively. Panels (c, f) display global composites of OND anomalies in sea surface temperature and precipitation (shading), 200 hPa geopotential (contours), and surface winds for positive events for periods P2 and P3, respectively. Attribution maps (b, e) indicate the relative contribution of each grid point in the predictor field to the forecasted value in the target region. The sum of the attribution values within each map equals the predicted anomaly in the corresponding index region (i.e., the sum of values in panel (b) matches the precipitation anomaly within the purple box in panel (a)). Panels (g, h) show the climatology of the SST and U200 for the period [1942-2007] as well as the differences from periods P2 and P3.