

<https://doi.org/10.1038/s43247-025-02207-2>

An interpretable machine learning model for seasonal precipitation forecasting

Check for updates

Enzo Pinheiro & Taha B. M. J. Ouarda

Seasonal climate forecasting is important for societal welfare, as it supports decision-makers in taking proactive steps to mitigate risks from adverse climate conditions or to take advantage of favorable ones. Here, we introduce TelNet, a sequence-to-sequence machine learning model for short-to-medium lead seasonal precipitation forecasting. The model takes past seasonal precipitation values and climate indices to predict an empirical precipitation distribution for every grid point of the target region for the next six overlapping seasons. TelNet has a simple encoder-decoder-head architecture, allowing the model to be trained with a limited amount of data, as is often the case in climate forecasting. Its deterministic and probabilistic performance is thoroughly evaluated and compared with state-of-the-art dynamical and deep learning models in a prominent region for seasonal forecasting studies due to its high climate predictability. The training, validation, and test sets are resampled multiple times to estimate the uncertainty associated with a small dataset. The results show that TelNet ranks among the most accurate and calibrated models across multiple initialization months and lead times, especially during the rainy season when the predictable signal is strongest. Moreover, the model allows instance- and lead-wise forecast interpretation through its variable selection weights.

Recent developments in sequence-to-sequence (seq2seq) machine learning models led to increased machine learning-based weather prediction (MLWP) models. The Convolutional Long-Short Term Memory (ConvLSTM) network was one of the first seq2seq models employed for precipitation nowcasting¹. The model successfully captured the spatio-temporal patterns of its training dataset, outperforming previous state-of-the-art nowcasting models. Nevertheless, Long-Short Term Memory (LSTM) networks are limited in modeling very long sequences². The self-attention mechanism, part of the transformer model³, solved these issues and led to a new generation of MLWP. For instance, MetNet was developed based on axial self-attention mechanism⁴ for probabilistic precipitation forecasts up to 8 h lead time and at a 1 km spatial resolution⁵. ClimaX, a foundation model for weather and climate modeling, was built upon vision transformer⁶ and able to cover a great range of tasks such as nowcasting, weather and subseasonal-to-seasonal (S2S) forecasting, and climate projections⁷. GraphCast⁸ was recently introduced for deterministic medium-range weather forecasts. It was based on Graph Neural Networks (GNN)^{9,10}, a network suitable for learning complex physics, such as weather dynamics. GenCast¹¹, a diffusion model developed as an adaptation of GraphCast for probabilistic forecasting, also used transformer blocks in addition to GNNs. GraphCast and GenCast models outperformed the state-of-the-art European Centre for

Medium-Range Weather Forecasts (ECMWF) weather forecasting system in several cases.

Numerical weather prediction (NWP) models, also known as dynamical models, map the current state of the atmosphere to future states by deterministically solving a set of partial differential equations that model the atmospheric dynamics. The rapid growth of uncertainty in the initial atmospheric condition and models' imperfect representations^{12,13} result in a limited forecasting horizon. Additionally, it is essential to track how the initial condition uncertainty evolves, which is done through ensemble forecasting. Ensemble forecasting uses the Monte Carlo method to approximate a stochastic dynamic forecast by repeatedly sampling from the initial condition probability distribution and running it through the deterministic NWP model¹⁴. This process is highly time-consuming and computationally intensive.

In contrast, MLWP models take advantage of other stochastic forecasting approaches that are faster at inference time and only computationally intensive at training time. For instance, MetNet directly maps a set of initial conditions to a probability distribution of target weather conditions⁵. GenCast¹¹ features a diffusion model that iteratively refines a possible candidate state, initialized as pure noise, conditioned on the previous two atmospheric states. The process is repeated with different noise samples to generate an ensemble of forecasts. Another example is the

Ensemble of Artificial Neural Networks¹⁵ for seasonal precipitation forecasting trained with different subsets of the original training set generated through the Bagging algorithm¹⁶.

Most of the seq2seq MLWP models mentioned previously were implemented for weather forecasting. In terms of seasonal forecasting, there is a gap in the applicability of these models for two main reasons: (1) Deep Neural Network (DNN) models are usually highly complex with millions of parameters, requiring a big dataset to learn its patterns properly. Climate datasets have a limited number of samples, i.e., they have only a few samples per year, whereas weather datasets can have thousands of samples per year; (2) weather forecasting is an initial condition problem, i.e., the next state of the atmosphere is conditioned on the previous ones, which makes autoregressive models, such as seq2seq models, perfectly suitable for the task. Seasonal forecasting is a boundary condition problem in which the future climate state depends more on interactions of the atmosphere with its boundaries, such as land and ocean, than on its previous states. In this context, large-scale climate oscillations (for short, climate oscillations) are recurring patterns of variability of the atmosphere circulation or coupled atmosphere-ocean system. Climate oscillations trigger teleconnection patterns, i.e., energy transport and wave propagation in the atmosphere and ocean circulation, that result in remote climate anomalies across the globe^{17–19}. Thus, climate oscillations yield climate predictability to some regions due to their teleconnections, and using their information in regression models is a natural choice for seasonal forecasting^{20–24}.

In this study, we introduce TelNet, an interpretable seq2seq model for probabilistic seasonal precipitation forecasting that considers past precipitation information in an autoregressive manner and climate oscillation indices as covariates. To the best of the authors' knowledge, seq2seq models were not yet employed for probabilistic seasonal forecasting. The model architecture is simpler than other seq2seq models to compensate for the limited number of samples available (Fig. 1). Moreover, we use overlapping seasonal values as an augmentation strategy to maximize the number of training samples. As will be shown in the following section, TelNet's forecasting skill ranks among the most accurate and calibrated models across multiple initialization months and lead times in a prominent region for seasonal forecasting studies.

Results

Comparison of TelNet with baseline models

This section compares TelNet and six baseline models deterministic and probabilistic performances. Only the results of forecasts made for seasons

that match dynamical models forecasts and that do not overlap with the season of the input data are shown. For instance, if the latest seasonal state available is November–December–January, the model can issue a forecast in early February for the next six overlapping seasons, from December–January–February (DJF) to May–June–July (MJJ), but we only present results for the last three seasons (MAM, AMJ, MJJ).

The model is evaluated for the state of Ceara in northeastern Brazil. This region is one of the most predictable regions worldwide in terms of seasonal forecasting^{25,26}, where both empirical and dynamical models have high forecasting skills²⁷. The seasonal precipitation cycle in the state of Ceara has a well-defined rainy period from mid-December to mid-July, which is further divided into three seasons, i.e., the pre-season in DJF, main season in March–April–May (MAM) and post-season in JJA, each with different dynamical rainfall precursors. In the pre-season, the rainfall is closely related to upper-level cyclonic activity associated with the South American monsoon system²⁸. The main season accounts for over 70% of the total annual rainfall, and its main rain-bearing system is the Intertropical Convergence Zone (ITCZ). The ITCZ positioning during this season depends on the inter-hemispheric gradient of SST anomalies in the tropical Atlantic sector²⁹. Finally, post-season rainfall only accounts for about 7% of the total annual rainfall and it is closely related to easterly waves that form in western Africa, propagate through the Atlantic, and reach the South American coast³⁰. The rest of the year is mostly dry with low seasonal totals, such as 11 mm on average in September–October–November (SON). Therefore, accurate forecasts for the year's first half are of great importance for sectorial decision-making in the state.

The median Root Mean Squared Error Skill Score (RMSESS) and Ranked Probability Skill Score (RPSS) scorebars are shown in Fig. 2 and Fig. 3, respectively. TelNet outperforms the baseline models across several initialization months and lead times, with accuracy improvement exceeding 30% in some cases. However, the large error bars highlight substantial variability in the skill scores due to differences in the testing samples. This variability is more pronounced for RMSESS compared to RPSS. Moreover, RPSS tends to improve as the forecasts approach the rainy season, whereas most dynamical models outperform TelNet during the dry season. Subsequently, we will discuss the models' performance for short- and medium lead forecasting. We examine the lead-1 forecasts issued in February and the lead-3 forecasts issued in November, as these forecasts target the rainy season when the predictable signal is strongest throughout the year.

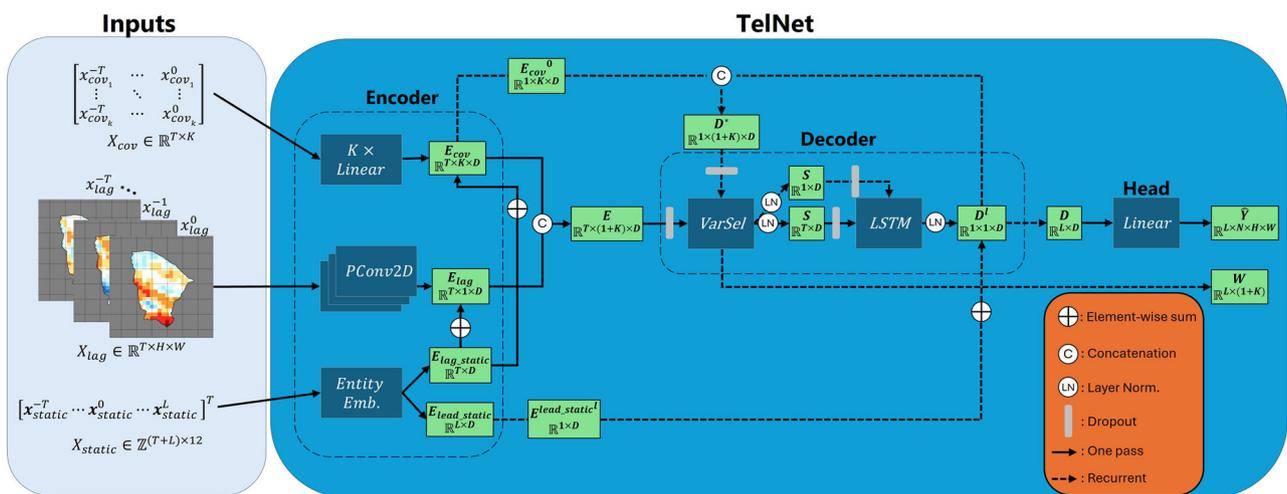


Fig. 1 | Schematics of the input dataset structure and model architecture. The model receives X_{cov} , X_{lag} , and X_{static} , along with a binary mask. Gray pixels in X_{lag} represent masked grid points. The input data set first goes through an encoder that maps the different arrays into model dimension D . The processed arrays go through a decoder that weights the features according to their importance and aggregates

them, followed by an LSTM. Finally, the prediction head maps the decoder output to an empirical distribution of standardized precipitation anomalies for every grid point. The model also outputs the variables selection weights, allowing the user to interpret the feature importance lead- and instance-wise.

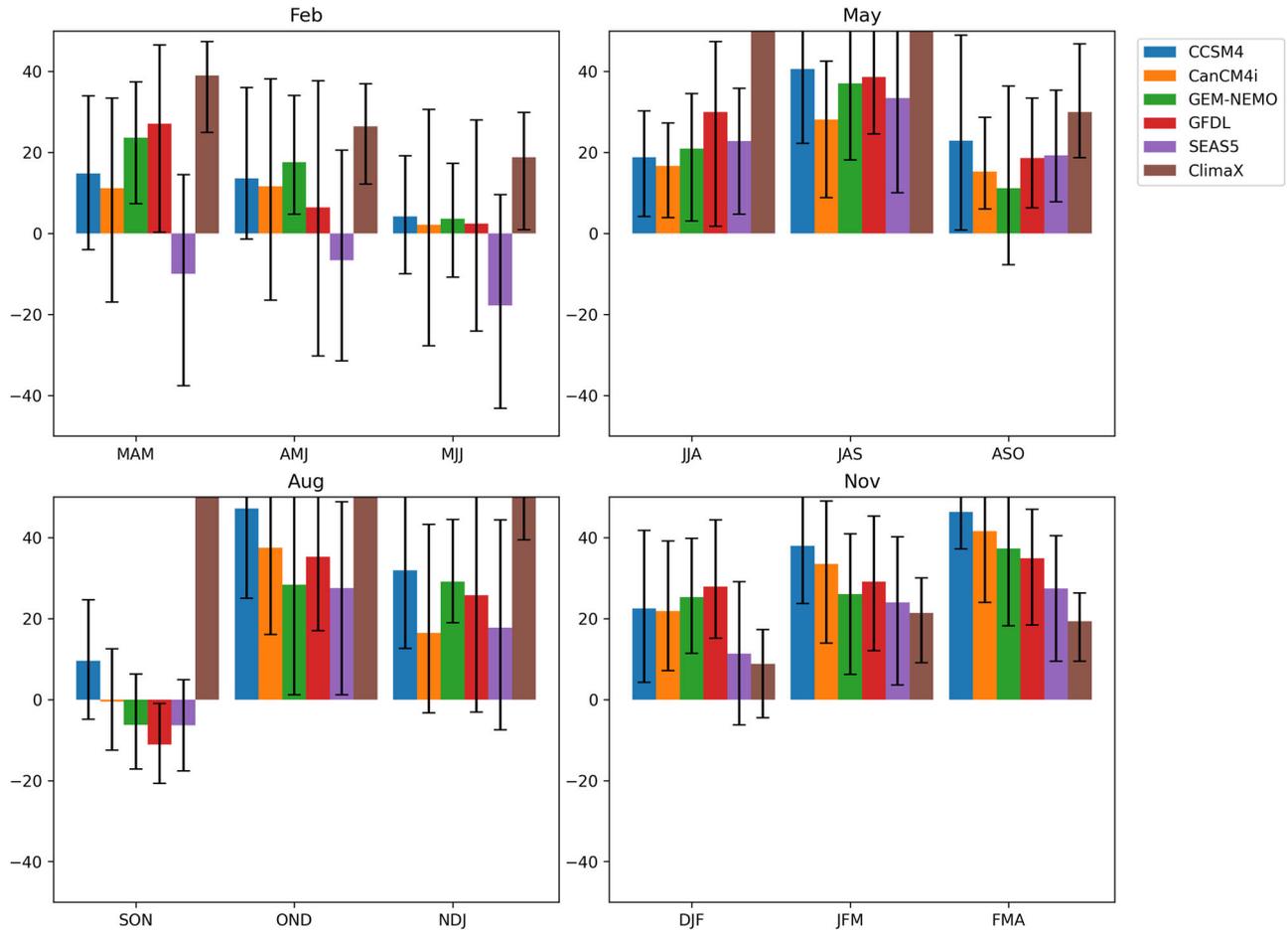


Fig. 2 | Median RMSESS [in %] scorebars computed on the test set across the last three lead times of February, May, August, and November initialization. Positive values indicate better performance of TelNet, and negative ones indicate better performance of baseline models. Error bars indicate the 90% confidence interval.

The median RMSESS (RPSS) for the MAM forecasts issued in February (lead 1), shown in Fig. 2 (Fig. 3), indicates that TelNet outperforms five (four) baseline models. The only exception is SEAS5 that yields up to a 17% (4%) improvement over TelNet in RMSESS (RPSS). The median skill scores for April-May-June (AMJ) forecasts are similar to those for MAM, although TelNet’s performance shows an overall decrease. For MJJ forecasts, the skill scores decrease further, with most models performing similarly to TelNet. However, SEAS5 outperforms TelNet by over 20% (15%) in deterministic (probabilistic) forecasting. In contrast, ClimaX underperforms TelNet by 20% in terms of RMSESS.

We further analyze the MAM forecasts issued in February by comparing the sample-mean spatial distribution of Root Mean Squared Error (RMSE) and Ranked Probability Score (RPS) across the models (Figs. 4 and 5). The RMSE patterns are generally consistent among the models, with smaller errors in the north than in the southern region. However, ClimaX exhibits the highest errors in the northernmost region, while GFDL yields the largest errors overall, especially in the south. SEAS5 presents the lowest RMSE in the north, whereas TelNet shows the lowest error in the southern region. The RPS maps show a similar pattern across the models, with smaller errors in the north than in the south. While most models perform well in the north, only TelNet and SEAS5 present small errors in the south. TelNet achieves the lowest sample-mean RPS.

The models rank histogram (Fig. 6) and reliability diagrams (Fig. 7) are analyzed next to assess other important attributes of probabilistic forecasting. Both diagrams’ statistics have been computed using area-aggregated statistics. Therefore, these results should be interpreted carefully given the high dimensionality of the verification problem and the small test set, which also shows substantial spatial correlation.

The rank histograms (Fig. 6) show that CCSM4 and GEM-NEMO are prone to under-dispersion, whereas GFDL and SEAS5 are slightly over-dispersed. CanCM4i ensemble members frequently overestimate observations, leading to a highly populated first bin. TelNet shows the flattest histogram but has large error bars, highlighting its sensitivity to the choice of the test set.

The reliability diagram (Fig. 7) reveals that TelNet and SEAS5 are calibrated for the Below Normal (BN) category (red curves), although both show an under-forecasting bias. The BN sharpness diagrams (red bars) of these two models are different, with SEAS5 featuring a positively skewed and TelNet a symmetric marginal distribution. The other four models show worse reliability and resolution than TelNet and SEAS5 for the BN category, featuring over-confidence when assigning high probabilities. For the Above Normal (AN) category (blue curves in Fig. 7), SEAS5 is the most reliable model, although it exhibits an over-forecasting bias. The other models display poor calibration. The large error bars in the AN calibration functions suggest high uncertainty, possibly due to the scarcity of wet events in the region over the last 20 years. Specifically, only three wet events occurred during Ceara’s rainy season within the test period, compared to seven dry events, including a multi-year drought from 2012 to 2016³¹.

To analyze the performance of TelNet for extreme events, Supplementary Fig. 1 shows the February-issued MAM rank histogram for events where the absolute observation value exceeds one standard deviation. The rank histogram shows that TelNet is prone to underestimation bias which is further evidenced in the MAM forecast examples from the test set of an extremely wet (Supplementary Fig. 2) and dry (Supplementary Fig. 3), respectively. Biased estimates of extreme values are a well-known issue of statistical models trained with loss functions such as the Mean Squared Error

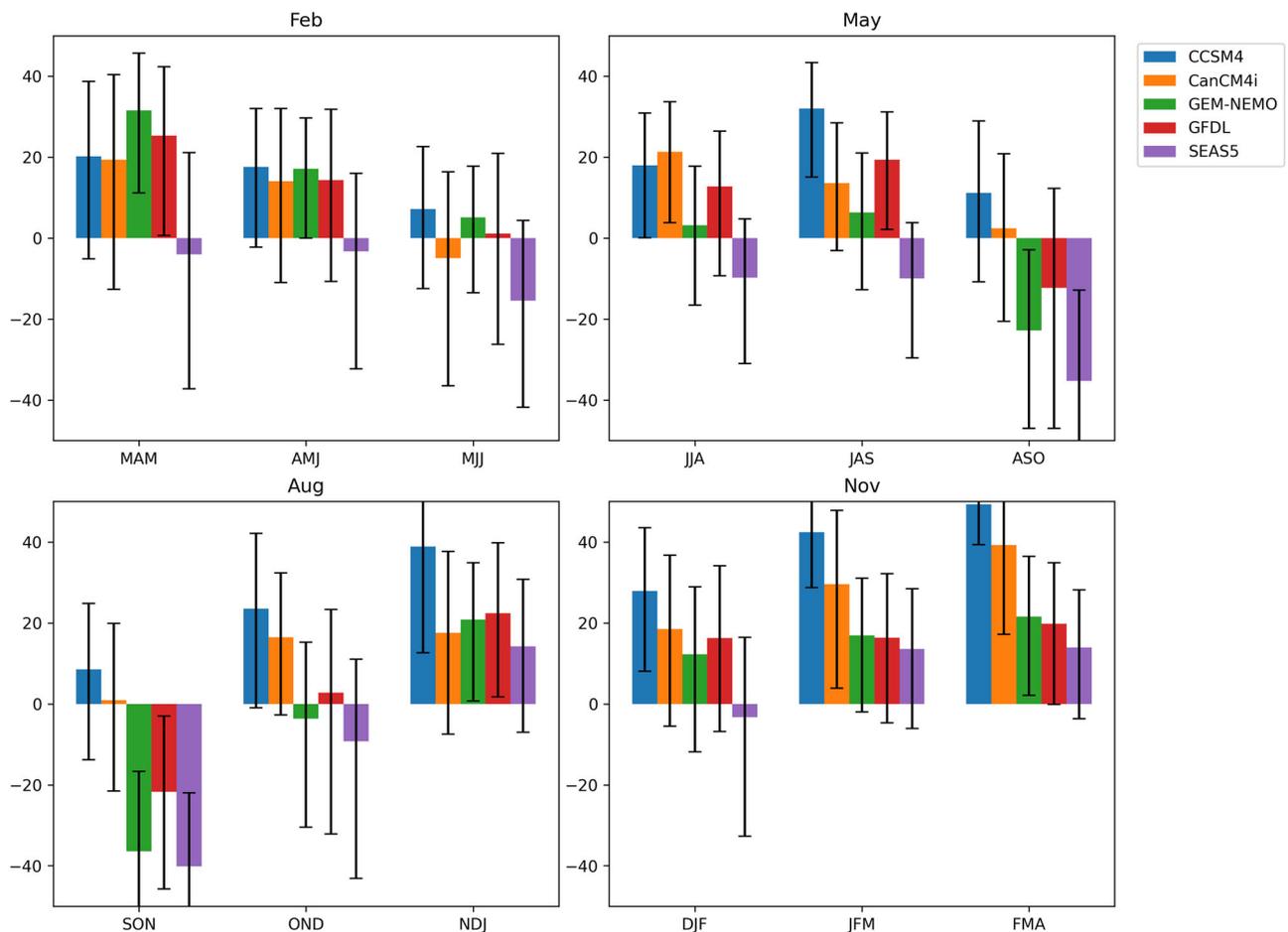


Fig. 3 | Median RPSS [in %] scorebars computed on the test set across the last three lead times of February, May, August, and November initialization. Positive values indicate better performance of TelNet, and negative ones indicate better performance of baseline models. Error bars indicate the 90% confidence interval.

or Median Absolute Error³². Overall, the dynamical model rank histograms for extreme events also exhibit underestimation bias, except for GFDL and SEAS5. Moreover, Supplementary Fig. 3 demonstrates that while TelNet often forecasts the correct most likely category, it exhibits underconfidence, i.e., ensemble members rarely converge on a single tercile. This underconfidence is reflected in the TelNet sharpness diagram (Fig. 7), which shows a low frequency of the highest probability bin (0.9–1.0) for extreme categories. Another study also reported a similar finding using an Ensemble of Artificial Neural Networks with a different ensemble-generation method¹⁵.

For FMA forecasts initialized in November (lead 3), the RMSESS (Fig. 2) and RPSS (Fig. 3) scorebars show that TelNet outperforms all baseline models. The sample-mean RMSE (Fig. 8) and RPS (Fig. 9) maps indicate worse performance among the models compared to the MAM forecasts issued in February, likely due to the longer lead time. TelNet consistently presents lower spatial errors across all regions of the state.

The rank histograms (Fig. 10) show that TelNet has the flattest histogram among the models, although with large error bars at the extremes. GEM-NEMO, GFDL, and SEAS5 are over-dispersed, while CCSM4 and CanCM4i are under-dispersed.

The reliability diagrams (Fig. 11) indicate that TelNet and GFDL are the most calibrated models for the BN category, while SEAS5 yields overconfident forecasts. All models feature low reliability for the AN category. Similar to the MAM forecasts issued in February, the limited number of wet events in the test period likely hinders a meaningful evaluation of the AN category.

Interpretability of the forecasts

An essential aspect of TelNet is that it allows users to identify features that play a prominent role in a specific forecast due to its variable

selection weights. Figure 12 shows the variable selection weights averaged over the test years for the last three leads of forecasts initialized in February, May, August, and November. On average, the variable selection weights show slight variation, suggesting that all indices contribute to the forecasts. However, differences among the indices are evident. For instance, the MAM forecasts issued in February tend to assign higher weights to the gradient of SST anomalies in the tropical Atlantic (ATL-SST), followed by the Oceanic Niño Index (ONI). This is further evidenced in the two forecast examples from the test set shown in Supplementary Figs. 2 and 3.

Lagged precipitation (YPRED) generally plays a minor role since autocorrelation is usually low for seasonal forecasting. However, in some instances, the model learns to use this information instead of climate indices, as indicated by the high 95th percentile of its weights.

Conclusions

In this study, we presented TelNet, an interpretable machine learning model for short-to-medium lead seasonal precipitation forecasting based on lagged climate indices (covariates) and precipitation. This is a simple model with three components: (1) an encoder that maps input data set with variables of different structure (spatial, temporal, and static variables) into the model hidden dimension; (2) a decoder that selects and aggregates the most prominent features instance- and lead-wise and process them through an LSTM layer to learn temporal relationships; and (3) a prediction head that maps the hidden features into an empirical distribution for every grid point of the target area.

In the variable and model selection phase, we sampled training and validation sets from 1941 to 2001 and pre-selected the most informative

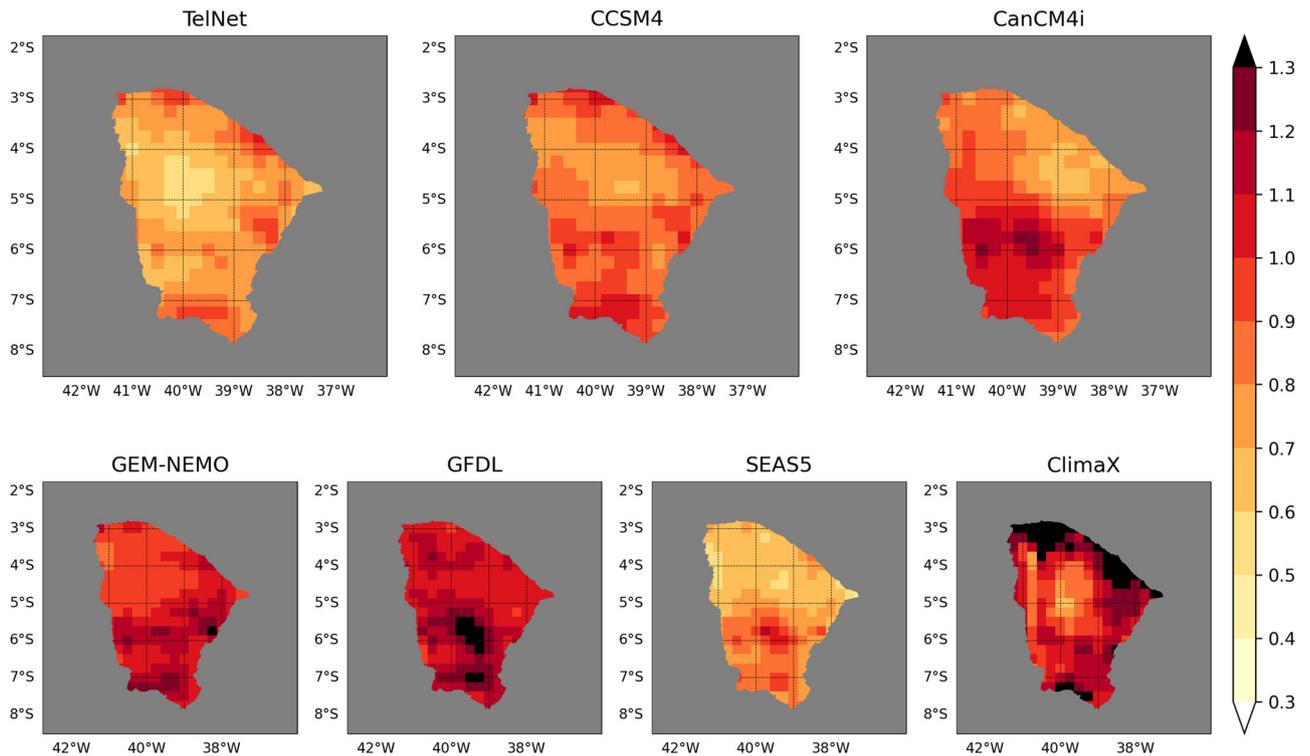


Fig. 4 | Mean RMSE maps of February-issued MAM (lead 1) forecasts of TelNet and baseline models computed on the test set. The colorbar is presented in standardized anomalies.

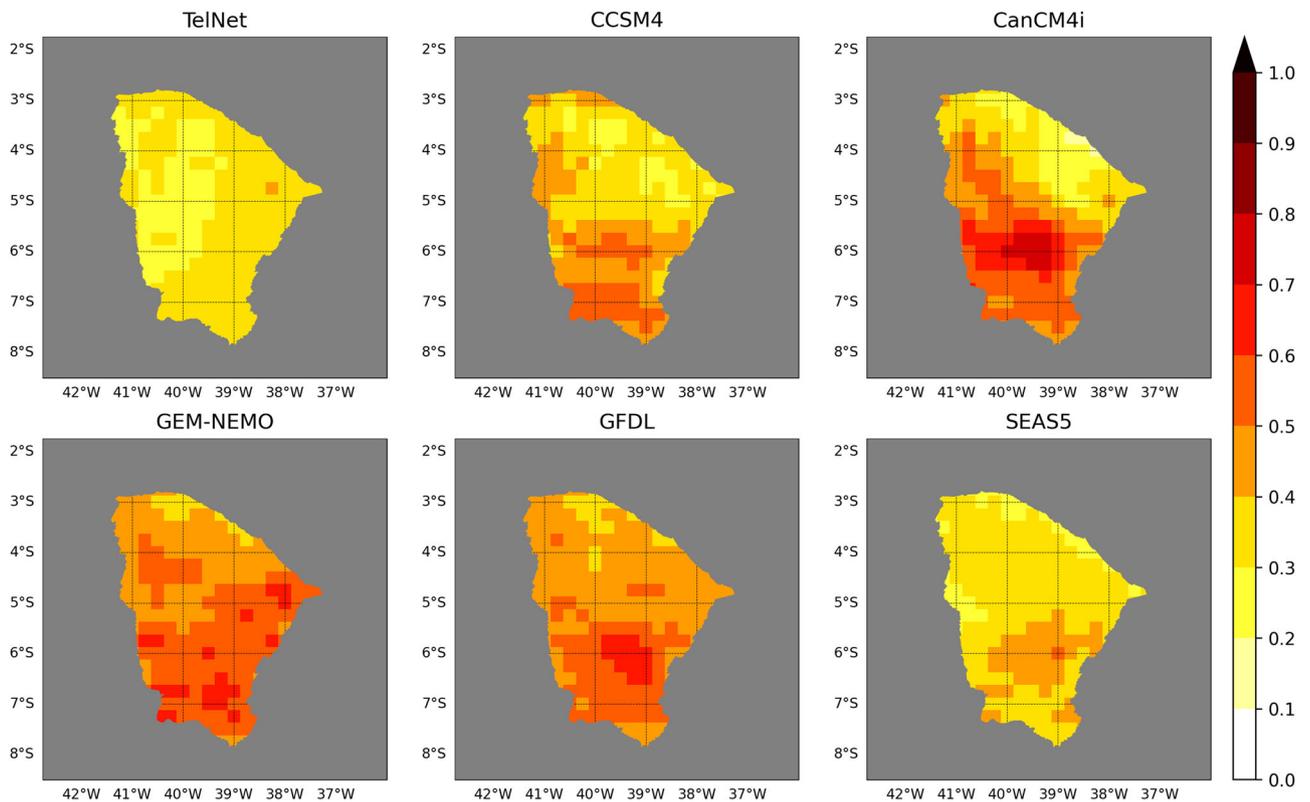


Fig. 5 | Mean RPS maps of February-issued MAM (lead 1) forecasts of TelNet and baseline models computed on the test set. The colorbar is unitless.

covariates for the predictand using the Partial Mutual Information method computed on the training set. TelNet was then trained using the pre-selected covariates, and the best architecture was chosen based on the average RPS computed on the validation set. In the testing phase, we retrained and

validated the model on the same period as the variable and model selection phase and tested on bootstrapped samples from 2003 to 2023. Both selection and testing phases were repeated 1000 times to account for the uncertainty associated with the limited sample size.

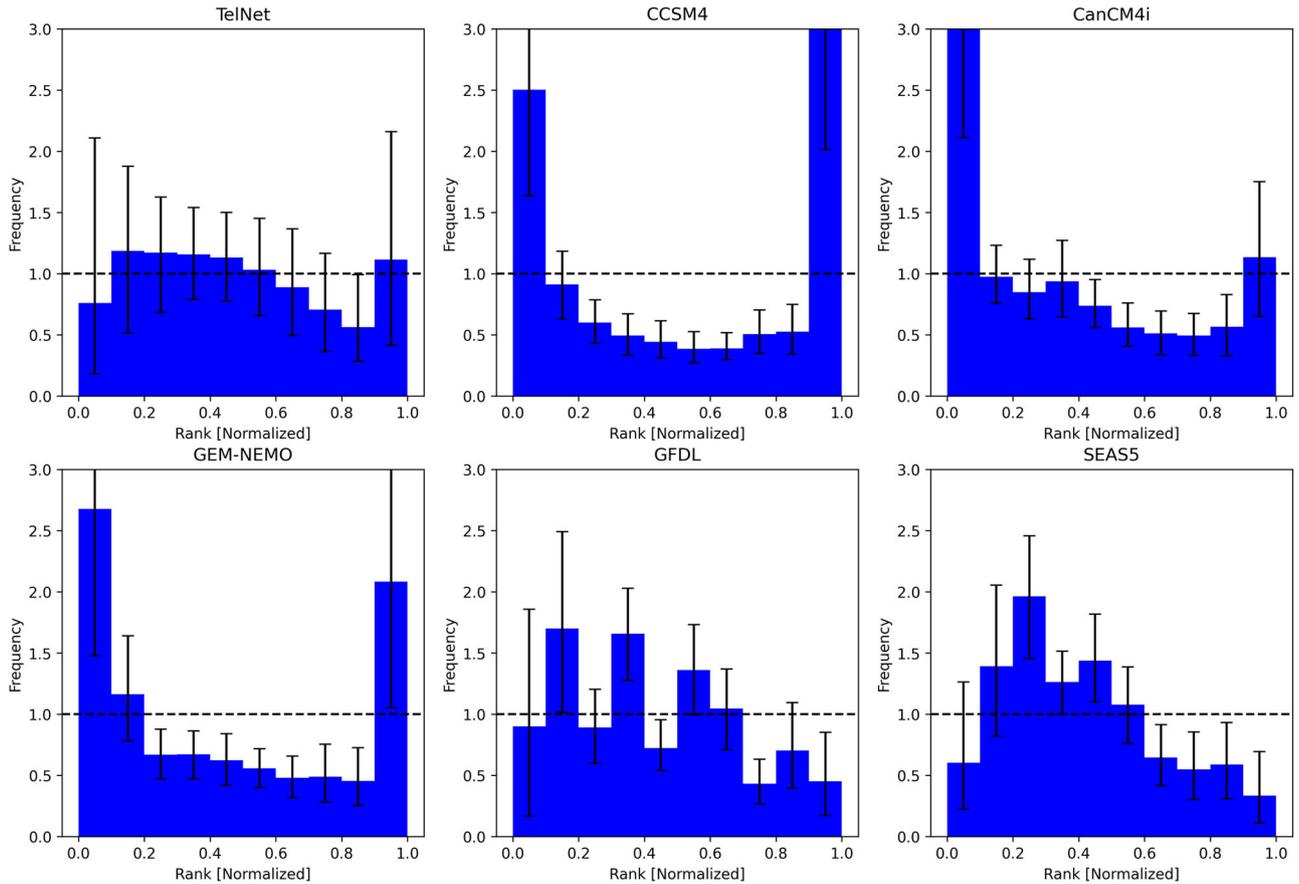


Fig. 6 | Median rank histogram of February-issued MAM (lead 1) forecasts of TelNet and baseline models computed on the test set. Normalization is done by dividing the observation rank by the number of ensemble members +1. Error bars indicate the 90% confidence interval.

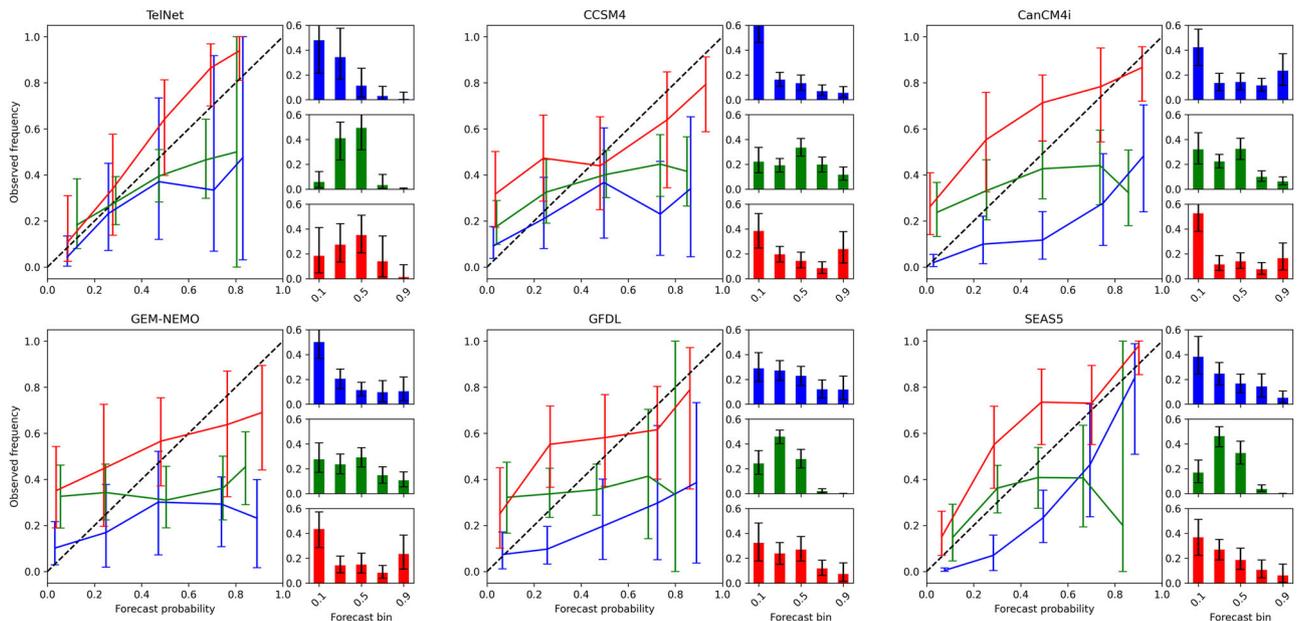


Fig. 7 | Reliability and sharpness diagrams of February-issued MAM (lead 1) forecasts of TelNet and baseline models computed on the test set. Blue lines and bars represent median probabilities for the above tercile, green the normal, and red the below. Error bars indicate the 90% confidence interval.

It is important to note that interleaved training and validation years were used. While we acknowledge that this approach evaluates the model’s ability to interpolate rather than predict future events in a dataset with autocorrelation, we selected this strategy to avoid having training, validation

and test periods too distant from each other. This choice was made to reduce the risk of evaluating the model under potentially different climate regimes caused by the non-stationarity of the climate system and ongoing climate change.

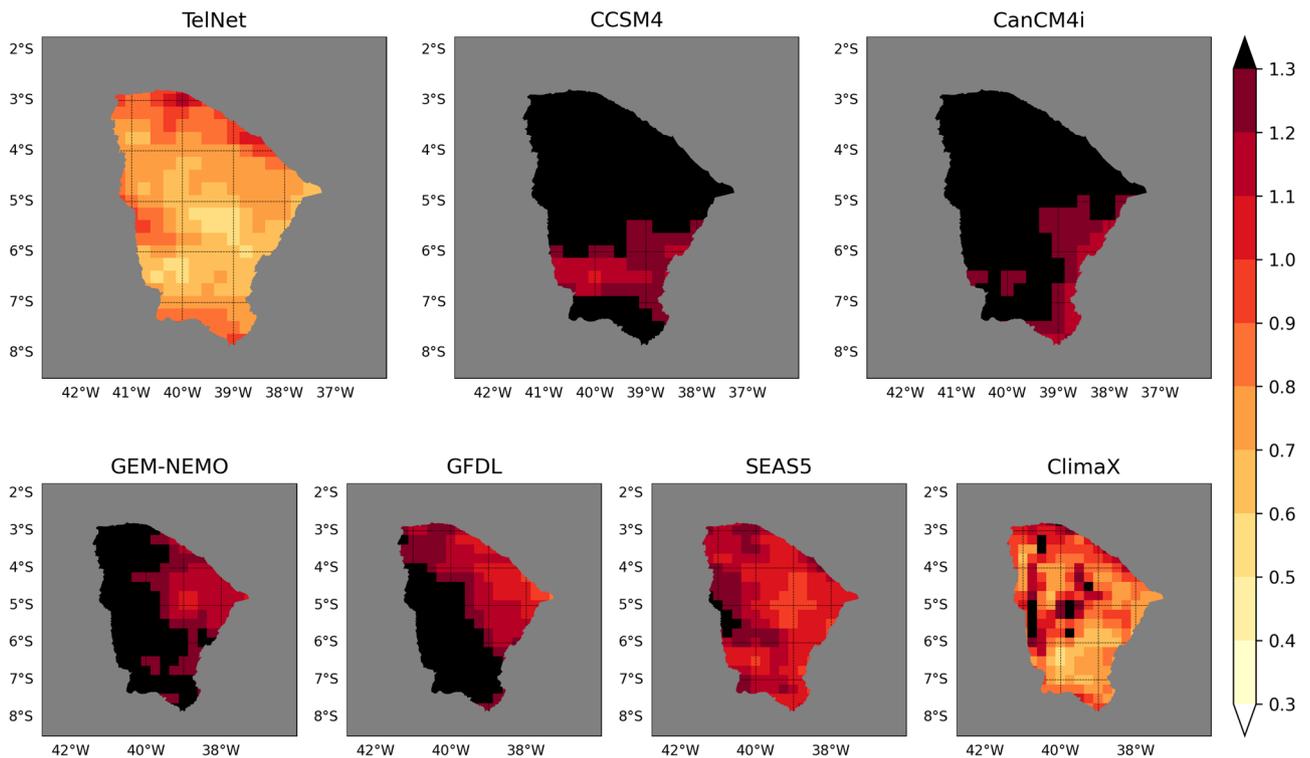


Fig. 8 | Mean RMSE maps of November-issued FMA (lead 3) forecasts of TelNet and baseline models computed on the test set. The colorbar is presented in standardized anomalies.

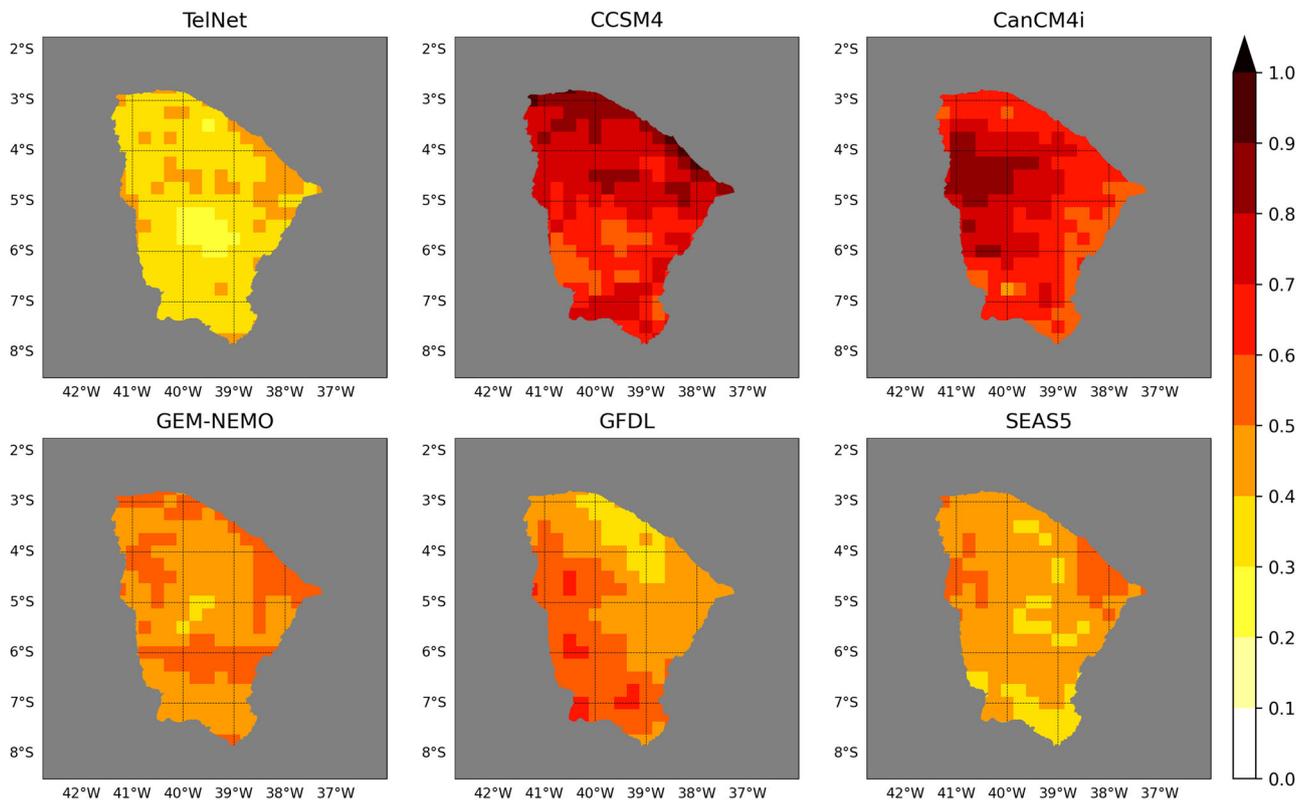


Fig. 9 | Mean RPS maps of November-issued FMA (lead 3) forecasts of TelNet and baseline models computed on the test set. The colorbar is unitless.

The model was evaluated in the state of Ceara, northeastern Brazil—a prominent region in seasonal forecasting studies due to its high predictability. Results demonstrated that TelNet outperformed several state-of-the-art baseline models for different initialization

months and lead times across the rainy season, when the predictable signal is strongest. However, TelNet performance decreased during the dry season and often underperformed dynamical models probabilistic skill.

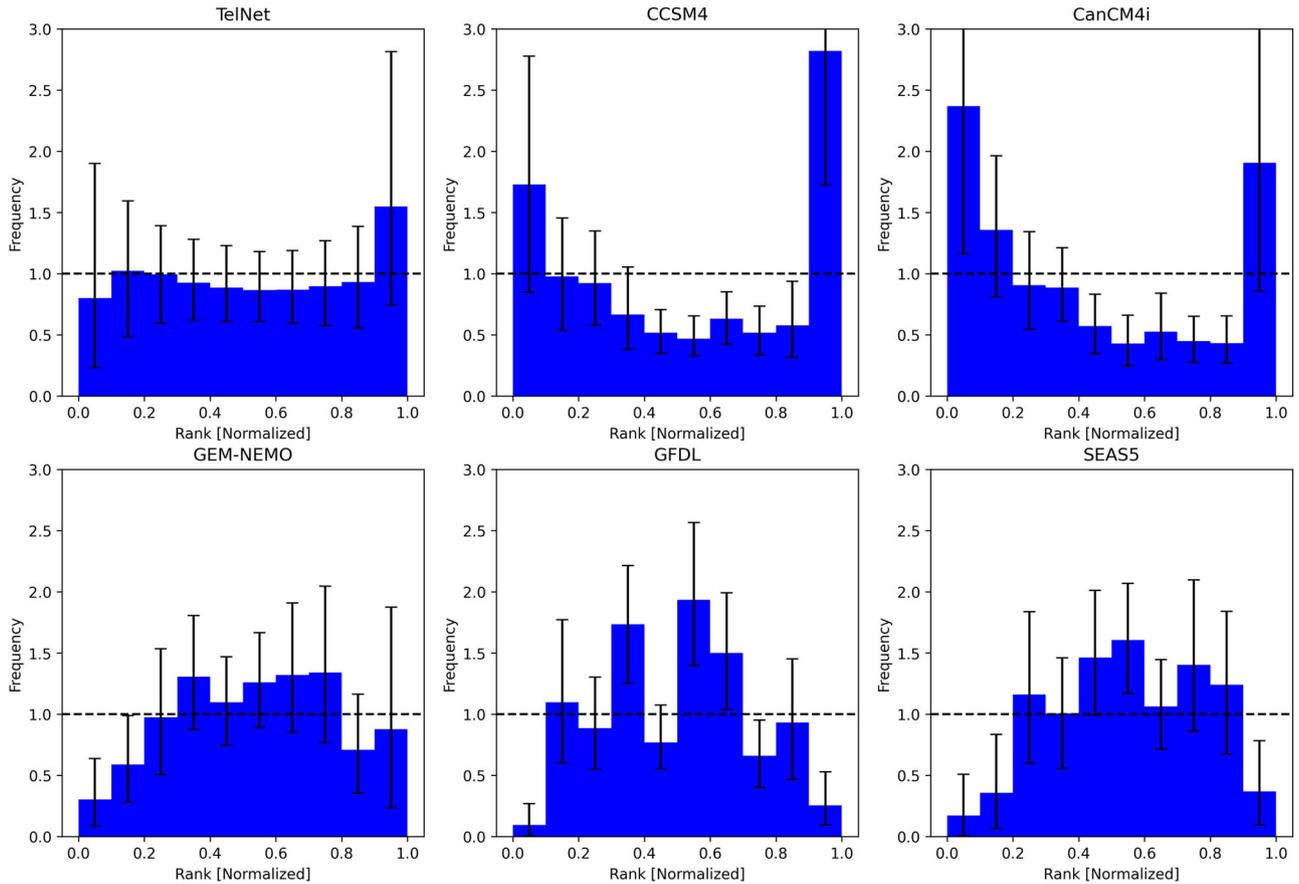


Fig. 10 | Median rank histogram of November-issued FMA (lead 3) forecasts of TelNet and baseline models computed on the test set. Normalization is done by dividing the observation rank by the number of ensemble members +1. Error bars indicate the 90% confidence interval.

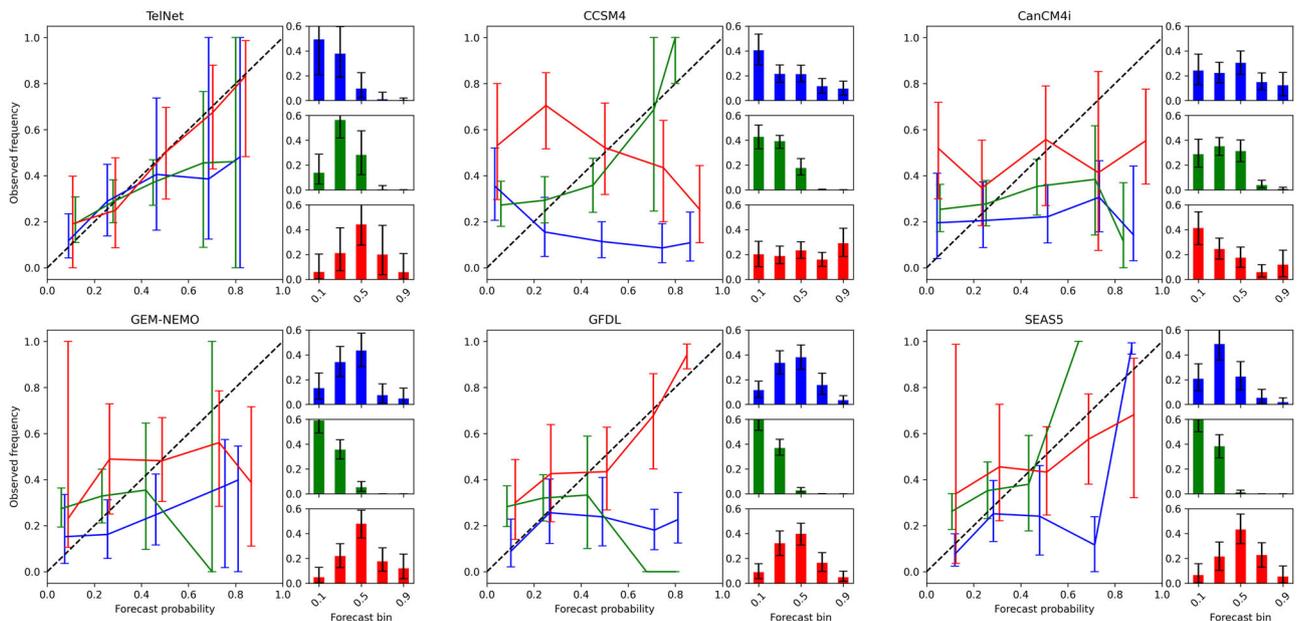


Fig. 11 | Reliability and sharpness diagrams of November-issued FMA (lead 3) forecasts of TelNet and baseline models computed on the test set. Blue lines and bars represent median probabilities for the above tercile, green the normal, and red the below. Error bars indicate the 90% confidence interval.

For the 1-month lead forecast of the MAM season, TelNet outperformed 5 (4) out of 6 (5) baseline models in deterministic (probabilistic) forecasting, with SEAS5 being the only model that consistently outperformed TelNet. For the 3-month lead FMA forecast, TelNet showed

better skill than all baseline models. Additional analysis of other probabilistic attributes showed that TelNet was among the most calibrated models for the BN category but exhibited low reliability for the AN category. However, it is important to note that during the testing period, only three wet events

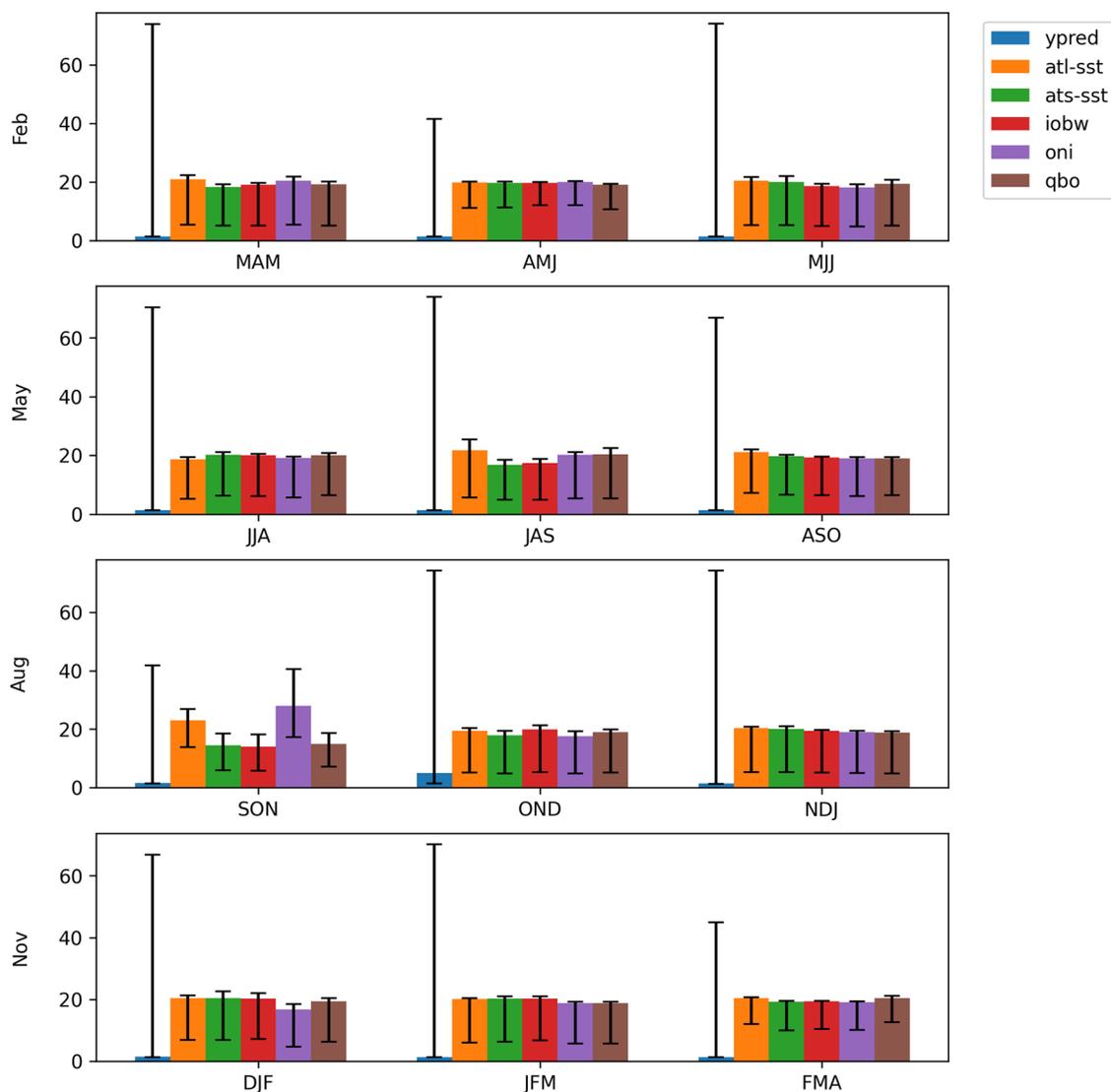


Fig. 12 | Average variable selection weights [in %] for the last three leads of forecasts initialized in February, May, August, and November of the test set. Bars represent the median value and error bars indicate the 90% confidence interval.

occurred in the rainy season, compared to seven dry events. This imbalance hinders a meaningful evaluation of the AN category. Future studies should consider alternative sampling strategies or testing methods to achieve a more balanced representation of events. Furthermore, while this study focused on Ceará, a region with high seasonal forecasting predictability, it is important to assess TelNet’s performance in other regions with different levels of predictability to better understand its broader applicability.

Another key feature of TelNet is its use of the Variable Selection Network of the Temporal Fusion Transformer³³, which assigns weights to features and allows users to interpret individual forecasts. Analysis of the variable selection weights indicated that the gradient of SST anomalies in the tropical Atlantic Ocean is often the most important feature followed by the Oceanic Niño Index, aligned with previous studies.

According to the results presented, TelNet is a forecasting model with high predictive skill, easy to implement, and computationally cheap at training and inference times. Nevertheless, TelNet is prone to underestimate extreme events and is underconfident in issuing the highest probabilities of extreme categories. Forecast postprocessing methods could be employed in future studies to improve those aspects. For instance, the ExEnsemble module is a parameter-free method that increases the variance of pixel values without altering the overall distribution, thereby producing more extreme predictions³². Moreover, the spatially distributed input X_{lag} was

encoded through a series of partial convolution operations that mapped the $H \times W$ grid points into model dimension D . This direct mapping works well for small regions with homogeneous precipitation patterns but could be limited for regions with high spatial variability. In this context, patch embedding could provide a better representation of spatial variability. This approach slices the spatial domain into patches, and each patch is embedded individually^{6,7}. Lastly, the model could also be applied to other forecasting tasks due to its flexibility in receiving past states of the target variable and covariates. For instance, sub-seasonal forecasting strongly relies on initial and boundary conditions, making TelNet a suitable choice for this task.

Data and methods

TelNet architecture

TelNet follows an approach similar to other MLWP models implemented for probabilistic forecasting⁵. The model directly predicts an empirical continuous distribution over the target seasonal climate conditions $Y^{1:L} = \hat{Y}$ for an L -month horizon conditioned on the T past seasonal states of the target variable $Y^{-T:0} = X_{lag}$, covariates $X^{-T:0} = X_{cov}$ and static variables X_{static}

$$p(Y^{1:L}|Y^{-T:0}, X^{-T:0}, X_{static}) = f_{\theta}(Y^{-T:0}, X^{-T:0}, X_{static}) \quad (1)$$

where $p(Y^{1:L}|Y^{-T:0}, X^{-T:0}, X_{static})$ is an empirical continuous distribution over the target $Y^{1:L}$ given the inputs $Y^{-T:0}, X^{-T:0}, X_{static}$ and $f_{\theta}(\cdot)$ is a neural network with learnable parameters θ .

The forecast model is made of three main components, as depicted in Fig. 1. The first layer encodes the inputs into the model dimension through linear and partial convolution operations³⁴. The decoder layer features a variable selection block³³ that selects and aggregates the candidate variables instance- and lead-wise, followed by a LSTM that learns the temporal relationship of the aggregated variable. Lastly, the decoder outputs go through a prediction head that forecasts an empirical distribution of N seasonal precipitation standardized anomalies for every grid point. Each component is detailed below.

Initially, each input type is encoded separately. The K covariates $\mathbf{X}_{cov} \in \mathbb{R}^{T \times K}$ with T past states are encoded individually through K linear layers, each with shared weights across the time dimension. The lagged state of the target variable $\mathbf{X}_{lag} \in \mathbb{R}^{T \times H \times W}$ with T past states, H latitudes, and W longitudes is encoded through a 3-layer sequence of partial 2D convolution (PConv2D)³⁴ with shared weights across the time dimension. The PConv2D layers were employed instead of regular Conv2D to allow users to train the model on a specific part of the H x W domain. For instance, one could train the model in areas with irregular shapes (such as a municipality or a watershed), by passing a binary mask where 1 represents the grid points of interest in the rectangular domain. The partial convolution operation at every location is expressed as

$$x' = \begin{cases} W^T(X \odot M) \frac{\text{sum}(1)}{\text{sum}(M)} + b, & \text{if } \text{sum}(M) > 0 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where \mathbf{X} is the pixel array, \mathbf{M} is the binary corresponding mask, \odot denotes element-wise multiplication, $\mathbf{1}$ has the same shape as \mathbf{M} but with all elements being 1. This layer also outputs an updated mask passed to the next PConv2D layer. A complete description of the PConv2D layer is found in the original paper³⁴.

We use the static variable $\mathbf{X}_{static} \in \mathbb{Z}^{(T+L) \times 12}$ to provide the model with information about the position of each past and future time step in a forecast. This variable consists of T vectors representing past time steps and L vectors representing the lead times. Each vector is one-hot encoded with 12 elements, where a value of 1 indicates the season corresponding to that position, and all other elements are set to 0. The static variable is embedded using entity embedding³⁵ and added to both encoded variables.

Finally, the encoded covariates and precipitation values are concatenated on the channel dimension.

The decoder comprises a Variable Selection Network³³ that provides an instance- and lead-wise variable selection and aggregation, followed by an LSTM that learns the temporal dependency of the aggregated variable.

The Variable Selection Network is a module designed to perform feature selection by assigning instance-wise weights to variables. This module was shown to be effective across various forecasting tasks, such as electricity consumption, traffic, and sales³³. In seasonal forecasting, instance-wise weights are particularly important because they capture the dynamic and complex relationships of covariates, which can create a wide variety of possible contexts. Additionally, the Variable Selection Network is based on a gating mechanism that allows the model to apply non-linear processing only where it is needed. This flexibility is especially beneficial for small datasets, such as climate datasets, where a simpler model can often yield better results. The Gated Residual Network (GRN) is defined as

$$GRN_{\omega}(a) = LayerNorm(a + GLU(\eta_1)) \quad (3)$$

$$\eta_1 = W_{1,\omega}\eta_2 + b_{1,\omega} \quad (4)$$

$$\eta_2 = ELU(W_{2,\omega}a + b_{2,\omega}) \quad (5)$$

$$GLU(\eta_1) = \sigma(W_{4,\omega}\eta_1 + b_{4,\omega}) \odot (W_{5,\omega}\eta_1 + b_{5,\omega}) \quad (6)$$

where ELU is the Exponential Linear Unit activation function, σ is the sigmoid activation function, $\eta_1 \in \mathbb{R}^D$ and $\eta_2 \in \mathbb{R}^D$ are intermediate layers, LayerNorm is the standard layer normalization³⁶, $\mathbf{W}_{(\cdot)} \in \mathbb{R}^{D \times D}$, $\mathbf{b}_{(\cdot)} \in \mathbb{R}^D$ are weights and biases, \odot is the element-wise multiplication, D is the hidden model dimension and ω is an index to denotes weight sharing. When $W_{2,\omega}a + b_{2,\omega} \gg 0$, the ELU activation would act as an identity function and when $W_{2,\omega}a + b_{2,\omega} \ll 0$, the ELU activation would generate a constant output, resulting in linear layer behavior³³.

For each time step, the variable selection network processes each of the 1 + K encoded variables through its own GRN, which are then weighted by their respective variable selection weights and combined:

$$S_t = \sum_{k=1}^{1+K} v_t^k \tilde{E}_t^k \quad (7)$$

$$\tilde{E}_t^k = GRN_{E(k)}(E_t^k) \quad (8)$$

$$v_t = Softmax(GRN_{v_t}(\cdot)) \quad (9)$$

where $\mathbf{t} = [\mathbf{E}_t^1, \dots, \mathbf{E}_t^{1+K}]^T$ is the flattened vector of the 1 + K encoded variables with $\mathbf{E}_t^k \in \mathbb{R}^D$ being the k-th encoded variable at time t, $\mathbf{v}_t \in \mathbb{R}^{1+K}$ is the variable selection weights, v_t^k is the k-th element of \mathbf{v}_t , GRN_{v_t} and $GRN_{E(k)}$ are the GRNs with weights shared across time dimension for the variable selection weights and for the k-th encoded variable, respectively. For a complete description of the Variable Selection Network, please refer to the original paper³³.

The combined variable $\mathbf{S} \in \mathbb{R}^{T \times D}$ goes through the LSTM layer, resulting in the first processed vector $\mathbf{D}^1 \in \mathbb{R}^D$, which is then added to its corresponding lead embedding $\mathbf{E}_{lead_static}^1 \in \mathbb{R}^D$. The resulting vector is concatenated in the channel dimension to $\mathbf{E}_{cov}^0 \in \mathbb{R}^{K \times D}$, representing the persistence of the latest available state of the covariates. This approach is common in seasonal forecasting models that do not prognostic the evolution of the boundary conditions³⁷. It often yields good results for short-term lead forecasts since boundary conditions represent slow-evolving components of the climate system^{38,39}. The resulting concatenated matrix $\mathbf{D}^{*1} \in \mathbb{R}^{1+K \times D}$ goes through the Variable Selection Network followed by the LSTM layer resulting in the next processed vector $\mathbf{D}^2 \in \mathbb{R}^D$. This process is repeated L times, where L is the model lead time. Each pass through the Variable Selection Network yields a different set of variable selection weights, allowing an instance- and lead-wise evaluation of variables relevance.

The resulting processed matrix $\mathbf{D} \in \mathbb{R}^{L \times D}$ goes through a linear layer that maps from the model dimension D to an empirical distribution of N seasonal precipitation standardized anomalies for every grid point of the target variable.

Datasets

The Extended Reconstructed Sea Surface Temperature (SST) version 5 (ERSST5) is used in the present study⁴⁰. This global gridded data set has a spatial resolution of 2° and spans from 1854 to the present. Monthly atmospheric variables from ERA5 reanalysis are also employed⁴¹. This is also a global gridded data set with a spatial resolution of 0.25° that covers the period from 1940 to the present. All variables have been linearly detrended before the analysis.

TelNet performance is compared with state-of-the-art numerical models from Copernicus Climate Change Service (C3S) and the North American Multi-Model Ensemble (NMME) project⁴². We use 1982–2021 monthly precipitation forecasts from the latest version of ECMWF seasonal forecasting system (SEAS5) and the latest version of the

NMME (NMME4) project⁴³. The monthly values are converted to total seasonal precipitation. Systematic bias in the mean and variance of dynamical models are corrected by subtracting forecasts from the model's long-term mean and then dividing by its long-term standard deviation. The bias correction is done independently for each model, target season and lead time.

Among the seq2seq MLWP models, only ClimaX⁷ is designed for long-range forecasts. Since the model produces point forecasts, we could only compare it to TelNet in terms of deterministic forecasting. Following the procedure for S2S predictions outlined in the original paper, we trained ClimaX from scratch in its default global forecasting configuration using seasonal values from ERA5 dataset, regrided to a 1.40625° resolution. The input variables are seasonally averaged values of the original input variables (as listed in Table 9 of the original paper⁷) and the output variable is total seasonal precipitation. The model is trained for forecasts from 1- to 3-month lead time on data from 1941 to 1993 and validated using data from 1994 to 2002.

The output of each baseline model is bilinearly interpolated onto ERA5 precipitation data set grid resolution for forecast verification.

Data preprocessing

Climate indices are computed using SST and atmospheric variables from ERSST5 and ERA5, respectively. Even though most indices are available online for download, their time series usually starts in 1948. To completely use ERA5 time availability, we decided to recompute those indices following methods well-established in the literature. The complete list of computed climate indices with their respective references is provided in Supplementary Table 1. Moreover, the Pearson correlation of the computed indices with their counterpart from NOAA's climate indices website (<https://psl.noaa.gov/data/climateindices/list/>) for their common period is also provided (Supplementary Table 1). All climate indices were normalized to unit variance.

ERA5 total monthly precipitation is converted to total seasonal values. Subsequently, season- and point-wise long-term mean, standard deviation, and terciles are computed from part of the training set covering 1971–2020. Standardized anomalies of the training, validation, and test sets are computed based on these long-term statistics. For forecast evaluation purposes, validation, and test sets total precipitation values are also categorized as AN, near normal (NN), and BN based on the computed terciles.

Variable selection module

The pool of candidate covariates (Supplementary Table 1) encompasses climate indices that might be relevant or not to the interannual precipitation variability of the target region. Therefore, a variable selection method is necessary to pre-select the most important climate indices for the region. Moreover, it is well established in the literature that teleconnections are nonlinear^{44,45}, requiring a suitable method for the task.

The Partial mutual information (PMI) method, a stepwise variable selection approach, is based on the mutual information (MI) criterion⁴⁶. This criterion measures the shared entropy between two random variables X and Y as follows

$$MI(X; Y) = \frac{1}{n} \sum_{i=1}^n \ln \left[\frac{f_{X,Y}(x_i, y_i)}{f_X(x_i)f_Y(y_i)} \right] \quad (10)$$

The PMI method uses a forward selection algorithm to select useful inputs from a candidate set C with $\{j = 1, \dots, J\}$ variables based on the maximum PMI between a given candidate input C_j and the output Z , controlling for the effects of the inputs that have already been selected, S . For the initial selection, S is an empty set, and the first input is selected as the one having maximum MI (Eq. 10) between the candidates in C and the output Z . In our study, C is the set of climate indices lagging four overlapping seasons behind Z , which is computed as the area-average standardized seasonal precipitation anomalies from the training set.

Next, the residuals w and v are computed as

$$w = C_j - \hat{C}_j(S) \quad (11)$$

$$v = Z - \hat{Z}(S) \quad (12)$$

where $\hat{C}_j(S)$ and $\hat{Z}(S)$ are estimators of C_j and Z , respectively. Differently from other papers that employed kernel density estimators^{46,47}, we used a simple 2-layer multi-layer perceptron of the form

$$\hat{C}_j(S) = W_2 \omega + b_2 \quad (13)$$

$$\omega = ReLU(W_1 S + b_1) \quad (14)$$

where $ReLU$ is the Rectified Linear Unit activation function, $S \in \mathbb{R}^{d_1}$ is the selected set with d_1 features, $\omega \in \mathbb{R}^{d_2}$ is an intermediate layer with $d_2 = 2d_1$ hidden features, $W_1 \in \mathbb{R}^{d_1 \times d_2}$, $b_1 \in \mathbb{R}^{d_2}$, $W_2 \in \mathbb{R}^{d_2 \times 1}$, $b_2 \in \mathbb{R}$ are the weights and biases. The estimator $\hat{C}_j(S)$ is fitted through backpropagation by minimizing the mean squared error. The estimator $\hat{Z}(S)$ can be similarly constructed.

Following, the next input C_j is selected as the one having maximum shared entropy with Z that has not been accounted for in S , estimated as

$$PMI(C; Z|S) = MI(w; v) \quad (15)$$

which is computed through Eq. 10. The selection process is repeated until all the candidate inputs in C are tested. The final set S is a PMI-ranked version of the initial set C .

Forecast verification

In seasonal forecasting, it is recommended that forecasts be issued in terms of the probability of equiprobable categories to reflect their uncertainties⁴⁸. These categories are usually three (AN, NN, BN) but there can also be more. We computed the probabilistic forecasts by counting the number of members that fall in each of the three equiprobable categories and dividing them by the total number of members N . The terciles for each season and grid point were previously computed based on the training set (Data preprocessing). Deterministic forecasts are computed as the ensemble mean.

The deterministic accuracy is measured as RMSE

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\bar{y}_i - o_i)^2} \quad (16)$$

where \bar{y}_i, o_i are the i -th of the n pairs of ensemble average and observation.

The probabilistic performance is measured through the RPS, which is an evaluation metric for multicategory events defined as

$$RPS = \frac{1}{n} \sum_{i=1}^n \sum_{m=1}^J \left[\left(\sum_{j=1}^m y_j \right) - \left(\sum_{j=1}^m o_j \right) \right]^2 \quad (17)$$

where y_j and o_j are the forecast and observation pair for the j -th category.

For a straightforward comparison with baseline models, we also used the RMSESS and the RPSS, defined as

$$RMSESS = \left(1 - \frac{RMSE_{TelNet}}{RMSE_{BL}} \right) \times 100 \quad (18)$$

$$RPSS = \left(1 - \frac{RPS_{TelNet}}{RPS_{BL}} \right) \times 100 \quad (19)$$

where the subscription "TelNet" represents the metrics computed for our model and the subscription "BL" for a given baseline model. The skill scores range from -100% to 100% , where positive values indicate better performance of TelNet and negative values the opposite. The spatial average

RMSESS and RPSS for specific initialization months and lead times are presented as scorebars.

The rank histogram is used to assess whether the forecast ensembles and the observation being forecasted are exchangeable. More precisely, the ensemble members and the corresponding observation are concatenated and ranked for each lead time and grid cell. Following, the collection of observation ranks is plotted in a histogram. If the observation and ensemble members have been drawn from the same distribution, then the rank of the observation is likely to be in any of the $N + 1$ ranks, resulting in a flat histogram. Observation values often ranked towards the boundaries of the histogram (U-shaped) reflect an under-dispersed ensemble, i.e., ensemble members are too similar to each other but different from the observation. Over-dispersion can be diagnosed by a rank histogram with relative frequencies centered in the middle ranks (∩-shaped), i.e., ensemble members frequently ranging beyond the observation value. Moreover, peaks on either extremity indicate over- or under-forecasting bias¹⁴.

Reliability and sharpness diagrams are also employed to assess three important aspects of probabilistic forecasts: reliability, resolution, and sharpness. Reliability measures the consistency between the forecast probabilities and the relative frequency of the observed outcomes. Resolution quantifies the degree to which the observed outcomes change as the forecasts change. Sharpness expresses how often each forecast probability is issued⁴⁹.

This study’s reliability and sharpness diagrams are based on a binning of 5 forecast probabilities over the whole geographic domain (area aggregated). The rank histogram is also constructed using area aggregated observation ranks.

Model selection

We start the analysis by performing variable and model selection through a grid search of 288 hyperparameter combinations. The training set consists of 51 years, resulting in a total of 612 overlapping seasons, and the validation set includes 10 years (120 overlapping seasons) randomly sampled from 1941 to 2001. To prevent any information leakage from the validation to the training set, samples that share the same seasonal value in different positions along the time dimension are removed from the training set.

Subsequently, training and validation sets are preprocessed (as described in “Data preprocessing”) and climate indices are ranked according to their PMI scores computed on the training set (Variable selection module). TelNet is then trained with each hyperparameter combination and evaluated on the validation set. The best model configuration was selected as the one that minimized the validation set average RPS (Eq. 17) taken over the valid grid points and all lead times.

As summarized in Table 1, the number of features in X_{cov} and the time dimension of X_{cov} and X_{lag} were also defined through grid search. The selected features correspond to the top K climate indices according to their PMI, ensuring that the model is built on the most relevant climate indices for the region.

This entire process is repeated 1000 times, and the final model configuration, along with the top K climate indices, is chosen as the one that most frequently appears as the best-performing model. The variable and

model selection entire procedure took about 3 days on a 40GB NVIDIA A100 GPU.

Model testing

Following the approach used for variable and model selection, 51 years from 1941 to 2001 are randomly sampled for the training set, and 10 years are assigned to the validation set, with the latter used for early stopping. The test set is created by bootstrapping years from 2003 to 2023. The training, validation and test sets are preprocessed as described in “Data preprocessing”, and the baseline models standardized anomalies and category probabilities are computed using a leave-one-out approach across all available years (Datasets). TelNet is then trained with the selected covariates and architecture (Model selection), and all models are evaluated on the test set.

To account for the uncertainty associated with a small dataset, the sampling, training, and evaluation procedures are repeated 1000 times. Once the process is complete, confidence intervals for these statistics are constructed independently for each model, season, and lead time, based on the 5th and 95th percentiles of the results.

The entire procedure required ~6 h on a 40GB NVIDIA A100 GPU.

Loss function

The training objective is based on the Continuous Ranked Probability Score (CRPS)⁵⁰. The CRPS is a strictly proper and negatively oriented score that compares the predicted $F(y)$ with the target variable ground truth translated into a step function $F_0(y)$:

$$CRPS = \int_{-\infty}^{\infty} [F(y) - F_0(y)]^2 dy \quad (20)$$

where

$$F_0(y) = \begin{cases} 0, & y < \text{observed value} \\ 1, & y \geq \text{observed value} \end{cases} \quad (21)$$

Its analytical form to estimate the expected CRPS of an empirical distribution of a finite ensemble of size N for a single instance is expressed as

$$\widehat{crps}(X, y) = \frac{1}{N} \sum_{n=1}^N |x^n - y| - \frac{1}{2N^2} \sum_{n,n'=1}^N |x^n - x^{n'}| \quad (22)$$

where $X \in \mathbb{R}^N$ and y is a scalar. Our loss function is defined as

$$Loss = \frac{1}{B \times L \times P} \sum_{b=1}^B \sum_{l=1}^L \sum_{p=1}^P w_p \widehat{crps}(\hat{Y}_{b,i,p}, y_{b,i,p}) \quad (23)$$

where B is the batch size, L is the number of leads, P is the number of valid (unmasked) grid points, and w_p is the latitude weighting factor of the p -th grid point, computed as

$$w_p = \frac{\cos(lat(p))}{\frac{1}{H} \sum_{h=1}^H \cos(lat(h))} \quad (24)$$

where H is the total number of latitudes. We used a Pyro⁵¹ CRPS implementation with an $N \log(N)$ complexity instead of the original algorithm with an N^2 complexity.

Data availability

ERSSTv5 is available at <https://www.ncei.noaa.gov/products/extended-reconstructed-sst>; ERA5 is available at <https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels-monthly-means?tab=overview>; The North American Multimodel Ensemble forecasts are available at <https://iridl.ldeo.columbia.edu/SOURCES/.Models/.NMME/>; SEAS5 forecasts are available at <https://cds.climate.copernicus.eu/>

Table 1 | Search matrix showing the six hyperparameters tuned for the model

Hyperparameter name	Values
Hidden dimension	128, 512, 1024
Dropout	0, 0.05, 0.25
Epochs	10 , 25
Learning rate	10^{-3} , 10^{-4}
# Features	2, 3, 4, 5
# Time steps	1 , 2

A total of 288 combinations were tested, and the hyperparameter set that minimized the average RPS computed on the validation set is highlighted in bold.

cdsapp#!/dataset/seasonal-monthly-single-levels?tab=overview. NOAA's climate indices repository: <https://psl.noaa.gov/data/climateindices/list/>.

Code availability

All codes necessary for reproducibility of the results are available at <https://github.com/enzopinheiro/telnet>.

Received: 29 August 2024; Accepted: 11 March 2025;

Published online: 21 March 2025

References

- Shi, X. et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Adv Neural Inf Process Syst* **2015-Janua**, 802–810 (2015).
- Hochreiter, S., Bengio, Y., Frasconi, P. & Schmidhuber, J. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. <https://doi.org/10.1109/9780470544037> (2001).
- Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017-Decem**, 5999–6009 (2017).
- Ho, J., Kalchbrenner, N., Weissenborn, D. & Salimans, T. Axial Attention in Multidimensional Transformers. 1–11 (2019).
- Sønderby, C. K. et al. MetNet: A Neural Weather Model for Precipitation Forecasting. 1–17 (2020).
- Dosovitskiy, A. et al. An image is worth 16×16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020).
- Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K. & Grover, A. ClimaX: A foundation model for weather and climate. 1–41 (2023).
- Lam, R. et al. Learning skillful medium-range global weather forecasting. *Science* **382**, 1416–1421 (2023).
- Alet, F. et al. Graph element networks: adaptive, structured computation and memory. In *Proc. 36th International Conference on Machine Learning, PMLR*, Vol. 97, 212–222 (2019).
- Pfaff, T., Fortunato, M., Sanchez-Gonzalez, A. & Battaglia, P. W. Learning mesh-based simulation with graph networks. *ICLR 2021 – 9th International Conference on Learning Representations* 1–18. arXiv preprint arXiv:2010.03409 (2021).
- Price, I. et al. Probabilistic weather forecasting with machine learning. *Nature* **637**, 84–90 (2025).
- Lorenz, E. N. The predictability of a flow which possesses many scales of motion. *Tellus A Dyn. Meteorol. Oceanogr.* **21**, 289 (1969).
- Hoskins, B. The potential for skill across the range of the seamless weather-climate prediction problem: A stimulus for our science. *Q. J. R. Meteorol. Soc.* **139**, 573–584 (2013).
- Wilks, D. S. *Statistical Methods in the Atmospheric Sciences*. 4th edition (Elsevier Science, 2019).
- Pinheiro, E. & Ouarda, T. B. M. J. Short-lead seasonal precipitation forecast in northeastern Brazil using an ensemble of artificial neural networks. *Sci. Rep.* **13**, 1–14 (2023).
- Breiman, L. Bagging predictors. *Mach. Learn* **24**, 123–140 (1996).
- An, S.-I., Wang, C. & Mechoso, C. R. Teleconnections in the Atmosphere. In *Interacting Climates of Ocean Basins* (ed. Mechoso, C.R.) Ch. 2, 54–82 (Cambridge University Press, 2021).
- Liu, Z. & Alexander, M. Atmospheric bridge, oceanic tunnel, and global climate teleconnections. *Rev. Geophys.* **45**, 1–34 (2007).
- Nagesh Kumar, D., Janga Reddy, M. & Maity, R. Regional rainfall forecasting using large scale climate teleconnections and artificial intelligence techniques. *J. Intell. Syst.* **16**, 307–322. (2007).
- Mekanik, F., Imteaz, M. A., Gato-Trinidad, S. & Elmahdi, A. Multiple regression and Artificial Neural Network for long-term rainfall forecasting using large scale climate modes. *J. Hydrol.* **503**, 11–21 (2013).
- Abbot, J. & Marohasy, J. Application of artificial neural networks to forecasting monthly rainfall one year in advance for locations within the murray darling basin, Australia. *Int. J. Sustain. Dev. Plan.* **12**, 1282–1298 (2017).
- Woldesellasse, H., Marpu, P. R. & Ouarda, T. B. M. J. Long-term forecasting of wind speed in the UAE using nonlinear canonical correlation analysis (NLCCA). *Arab. J. Geosci.* **13**, 962 (2020).
- Elshaboury, N., Elshourbagy, M., Al-Sakkaf, A. & Abdelkader, E. M. Rainfall forecasting in arid regions using an ensemble of artificial neural networks. *J. Phys. Conf. Ser.* **1900**, 012015 (2021).
- Coelho, C. A. S., Stephenson, D. B., Balmaseda, M., Doblas-Reyes, F. J. & van Oldenborgh, G. J. Toward an integrated seasonal forecasting system for South America. *J. Clim.* **19**, 3704–3721 (2006).
- Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P. & Rodrigues, L. R. L. Seasonal climate predictability and forecasting: status and prospects. *Wiley Interdiscip. Rev. Clim. Chang.* **4**, 245–268 (2013).
- Le, P. V. V. et al. Climate-driven changes in the predictability of seasonal precipitation. *Nat. Commun.* **14**, 1–10 (2023).
- Hastenrath, S., Sun, L. & Moura, A. D. Climate prediction for Brazil's Nordeste by empirical and numerical modeling methods. *Int. J. Climatol.* **90**, 921–926 (2009).
- Zhou, J. & Lau, K. M. Does a monsoon climate exist over South America? *J. Clim.* **11**, 1020–1040 (1998).
- Hastenrath, S. Circulation and teleconnection mechanisms of Northeast Brazil droughts. *Prog. Oceanogr.* **70**, 407–415 (2006).
- Giraldo-Cardenas, S., Arias, P. A., Vieira, S. C. & Zuluaga, M. D. Easterly waves and precipitation over northern South America and the Caribbean. *Int. J. Climatol.* **42**, 1483–1499 (2022).
- Funceme. Ceara's rainfall calendar. <https://chuvass.funceme.br/> (2024).
- Xu, W. et al. ExtremeCast: Boosting Extreme Value Prediction for Global Weather Forecast (2024).
- Lim, B., Arkk, S., Loeff, N. & Pfister, T. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *Int. J. Forecast.* **37**, 1748–1764 (2021).
- Liu, G. et al. Image inpainting for irregular holes using partial convolutions. In *Proc. Lecture Notes in Computer Science*, Vol. 11215, 89–105 (2018).
- Guo, C. & Berkhahn, F. Entity Embeddings of Categorical Variables. 1–9 (2016).
- Ba, J. L., Kiros, J. R. & Hinton, G. E. Layer Normalization. (2016).
- Goddard, L. & Mason, S. Sensitivity of seasonal climate forecasts to persisted SST anomalies. *Clim. Dyn.* **19**, 619–631 (2002).
- Latif, M. et al. A review of the predictability and prediction of ENSO. *J. Geophys. Res. Oceans* **103**, 14375–14393 (1998).
- Lee, T., Ouarda, T. B. M. J. & Seidou, O. Characterizing and forecasting climate indices using time series models. *Theor. Appl. Climatol.* **152**, 455–471 (2023).
- Huang, B. et al. Extended reconstructed Sea surface temperature, Version 5 (ERSSTv5): upgrades, validations, and intercomparisons. *J. Clim.* **30**, 8179–8205 (2017).
- Hersbach, H. et al. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146**, 1999–2049 (2020).
- Kirtman, B. P. et al. The North American multimodel ensemble: phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. *Bull. Am. Meteorol. Soc.* **95**, 585–601 (2014).
- Becker, E., Kirtman, B. P. & Pegion, K. Evolution of the North American Multi-Model Ensemble. *Geophys. Res. Lett.* **47**, 1–9 (2020).
- Hoerling, M. P., Kumar, A. & Zhong, M. El Niño, La Niña, and the nonlinearity of their teleconnections. *J. Clim.* **10**, 1769–1786 (1997).
- Hounsou-Gbo, A., Servain, J., das Vasconcelos Junior, F. C., Martins, E. S. P. R. & Araújo, M. Summer and winter Atlantic Niño: connections with ENSO and implications. *Clim. Dyn.* **55**, 2939–2956 (2020).
- Sharma, A. Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: part 1 — a strategy for system predictor identification. *J. Hydrol.* **239**, 232–239 (2000).
- May, R. J., Maier, H. R., Dandy, G. C. & Fernando, T. M. K. G. Non-linear variable selection for artificial neural networks using partial mutual information. *Environ. Model. Softw.* **23**, 1312–1326 (2008).

48. WMO. *WMO Guidance on Operational Practices for Objective Seasonal Forecasting*. ISBN: 978-92-63-11246-9 (World Meteorological Organization, 2019).
49. Jolliffe, I. T. & Stephenson, D. B. *Forecast Verification* (John Wiley and Sons, 2012). <https://doi.org/10.1002/9781119960003>.
50. Gneiting, T. & Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102**, 359–378 (2007).
51. Pyro-PPL. Pyro CRPS algorithm. Preprint at https://docs.pyro.ai/en/stable/ops.html#pyro.ops.stats.crps_empirical (2018).

Acknowledgements

The authors thank the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canada Research Chairs Program for funding this research. We also thank Copernicus Climate Change Service for providing the ERA5 and SEAS5 data sets. NMME project and data dissemination is supported by NOAA, NSF, NASA, and DOE. The NMME forecasts and data archive are created, updated, and maintained by NCEP, IRI, and NCAR personnel. Lastly, we thank the Digital Research Alliance of Canada for providing the computational resources to conduct this research.

Author contributions

E.P. conceived and designed the study, conducted the analysis, and wrote the main manuscript with input from T.B.M.J.O. All authors contributed to the reviewing, editing, and improvement of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s43247-025-02207-2>.

Correspondence and requests for materials should be addressed to Enzo Pinheiro.

Peer review information *Communications Earth & Environment* thanks Timothy Stockdale and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Rahim Barzegar and Alireza Bahadori. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025