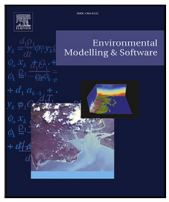
Environmental Modelling and Software 196 (2026) 106765

Contents lists available at ScienceDirect

Environmental Modelling and Software

journal homepage: www elsevier com/locate/envsoft



CRPS-Net: A novel framework for AI-assisted meteo-marine ensemble forecasting

Mattia Cavaiola [a], Daniele Lagomarsino-Oneto [a], Andrea Mazzino [b,c,*]

[a] CNR - National Research Council of Italy Institute of Marine Sciences Via S.Teresa S/N Pozzuolo di Lerici 19032 La Spezia Italy
[b] DICCA Department of Civil Chemical and Environmental Engineering Via Montallegro 1 Genova 16145 Italy
[c] INFN Istituto Nazionale di Fisica Nucleare Sezione di Genova Via Dodecaneso 33 Genova 16146 Italy

ARTICLE INFO

Keywords:
Ensemble forecasting
Deep learning
Hybrid modeling
Met-ocean modeling

ABSTRACT

This study introduces a novel non-parametric framework for generating ensemble forecasts directly from a single deterministic prediction. Unlike existing approaches that rely on pre-existing ensembles or impose a predefined parametric form on the error distribution (e.g., Gaussian), our method leverages a single deep neural network trained using the Continuous Ranked Probability Score (CRPS) as a loss function—without making any assumption on the shape of the predictive distribution. This non-parametric nature enables the model to flexibly learn complex, potentially multi-modal or skewed distributions that would be poorly captured by classical parametric methods. Applied to 24-hour wind forecasts and sea surface temperature time series, our approach yields ensemble outputs with improved probabilistic calibration compared to standard parametric techniques. Its ability to produce reliable probabilistic forecasts from a single deterministic input makes it a scalable and cost-effective alternative to traditional ensemble systems, particularly valuable in operational forecasting and long-range climate applications where computational constraints are critical.

1. Introduction

In meteorology and oceanography, ensemble forecasting has long been a cornerstone for capturing uncertainties inherent in numerical weather prediction (NWP) and marine modeling systems (Leith, 1974; Epstein, 1969). Traditional ensemble forecasting strategies, such as those used by the Ensemble Prediction System (EPS) of the European Centre for Medium-Range Weather Forecasts (ECMWF), rely on perturbations of initial conditions and model parameters to generate a set of possible future states of the atmosphere or ocean (Leutbecher et al., 2008; Buizza et al., 1999). These systems offer significant benefits in terms of probabilistic forecasting, yet they come with substantial computational costs due to the necessity of running multiple simulations with varied inputs (Houtekamer and Zhang, 2016; Palmer, 2019; Tuju et al., 2022).

In marine forecasting, ensemble systems are widely used for wave height, sea surface temperature (SST), and current predictions (Cavaleri et al., 2010; Mogensen and Bell, 2012). For example, multi-model ensembles have proven particularly useful for ocean reanalysis, where uncertainties in initial conditions and model dynamics are significant (Ferry et al., 2010; Stammer et al., 2016; Balmaseda et al., 2015). Recent advances in ensemble forecasting have significantly extended the prediction horizon for ocean mesoscale, providing greater predictability for up to 40 days, further enhancing ocean forecasting capabilities (Thoppil et al., 2021).

In the realm of climate modeling, traditional ensemble strategies are employed to capture the uncertainties of long-term climate predictions. Much like in NWP, climate ensembles rely on perturbing initial and boundary conditions to simulate a range of potential future states. This approach is pivotal for assessing climate sensitivity, understanding potential extremes, and evaluating the impacts of various forcings on climate projections (Palmer, 2019). For example, the Coupled Model Intercomparison Project (CMIP) uses multi-model ensembles to evaluate uncertainties in future climate scenarios, which are crucial for understanding variability and predictability over long timescales (Stouffer et al., 2017).

Although ensemble forecasting clearly brings about great improvements to meteo-marine modeling and climate studies, the computational burden associated with these systems limits their operational use, particularly when high-resolution, real-time and long-term forecasting is required.

Another important aspect, strictly connected with uncertainty estimation, is forecast calibration. Indeed, for some applications the output of either ensemble or single models require to be refined to increase

* Correspondence to: CNR - National Research Council of Italy, Institute of Marine Sciences, Via S.Teresa S/N, Pozzuolo di Lerici, 19032, La Spezia, Italy.
E-mail address: andrea.mazzino@unige.it (A. Mazzino).

https://doi.org/10.1016/j.envsoft.2025.106765
Received 21 December 2024; Received in revised form 4 September 2025; Accepted 1 November 2025
Available online 5 November 2025

performances on specific predictive tasks. A large body of research has been developed to tackle this problem by using statistical post-processing methods, such as Bayesian Model Averaging (BMA) (Raftery et al., 2005), Ensemble Model Output Statistics (EMOS) (Gneiting and Raftery, 2005) and the Ensemble Kalman Filter (EnKF) (Evensen, 2003). These techniques produce refined ensemble outputs by using statistical methods to improve forecast calibration and sharpness (Whitaker and Hamill, 2002).

Recently, hybrid data-driven models employing a multi-model ensemble approach have shown great promise in improving forecast accuracy and operational efficiency by combining traditional numerical weather prediction techniques with advanced machine learning methods, specifically targeting hydrological forecasting of variables like discharge and water levels (Hauswirth et al., 2023). With the advent of machine learning (ML) and deep learning (DL), AI-based strategies for ensemble forecasting have emerged, showing great potential in both post-processing and generation of ensemble forecasts (Zhang and Hong, 2019). AI methods have primarily been applied to calibrate pre-existing ensemble forecasts, correcting systematic biases and enhancing probabilistic accuracy (Rasp and Lerch, 2018; Scher and Messori, 2019). However, these approaches still rely on a pre-existing ensemble as input, limiting their potential in scenarios where ensemble generation is computationally prohibitive. The work by Lee et al. (2022) employs a Multi-Model Ensemble Skill Score (MMESS) approach, combining neural networks like CNNs and GRUs with statistical models to significantly boost forecast accuracy of next-day maximum air temperature. While these ensemble methods clearly improve prediction reliability, they also entail high computational costs due to the need to train multiple models.

Beside computational resources, there is a more fundamental limitation. Indeed, both statistical or ML models for generating ensemble forecasts, often require assuming a prior distribution for errors, e.g. Gaussian, but also gamma or logistic - which may not accurately capture the error structure in highly nonlinear dynamical systems like the atmosphere and oceans. The need to define *a priori* the error distribution can limit the flexibility and accuracy of these methods in representing the full range of uncertainties. Generating ensembles from deterministic forecasts has been already pursued also with classical statistical techniques like, for example, the Bayesian Joint Probability approach (Zhao et al., 2015; Samal et al., 2023), demonstrating the capability to go beyond the Gaussianity assumption. However, such methods are ultimately parametric, hence potentially incurring in fundamental limitations.

In summary, despite the significant advancements in both traditional and AI-enhanced ensemble forecasting, several critical gaps remain. Traditional ensemble systems, while effective, demand immense computational resources, which limits their feasibility to only certain types of real-time applications (Palmer et al., 2000). AI-based methods, although promising, have primarily focused on ensemble calibration rather than generation, thus still relying on expensive input ensembles. When not used for calibrating an ensemble of input models, these approaches often generate ensembles through multiple neural networks, leading to increased computational costs due to the need to perform multiple trainings. Statistical or AI-based approaches, when applied to deterministic inputs, reduce computational costs but typically require strong assumptions about the distribution of errors, which may not be appropriate in the context of highly nonlinear systems (Hamill and Whitaker, 2006; Gneiting, 2014).

Here we propose a new method for probabilistic calibrations of deterministic forecasts, where a non-parametric form of the Continuous Ranked Probability Score (CRPS) is implemented in a Deep Learning framework. Hence, our approach combines the sustainability of building upon deterministic forecasts, which has been already explored in previous works, with high flexibility in describing the probability distribution of predicted variables, which is a novel feature. It is worth underlining that predicted distributions not only depend on the considered application, but are also conditioned by inputs. Therefore, the persistence of any functional form across samples is not guaranteed, and any related assumption potentially affects performances.

Regarding sustainability, our approach significantly reduces computational overhead while retaining forecast skill. It thus appears particularly valuable both in operational settings, where computational resources are limited and quick probabilistic forecasts are essential, and also in climate settings.

The optimization of our neural network is done employing the Continuous Ranked Probability Score (CRPS) as a loss function. CRPS has previously been applied in ensemble post-processing (Gneiting and Raftery, 2005; Casciaro et al., 2022), and some recent studies have used it for AI-based ensemble calibration (Rasp and Lerch, 2018). However, these approaches, unlike our approach, assume a prior distribution for errors and require pre-existing ensembles as input.

The fact that our approach generates the ensemble through a single neural network eliminates the need to train multiple networks, significantly reducing computational costs and simplifying the implementation process. Additionally, by utilizing CRPS as a loss function without assuming Gaussianity, we extend the applicability of this method to a broader range of forecasting scenarios.

The new ensemble strategy is applied to two significant cases in atmospheric and oceanic sciences: (i) the 24-hour wind forecast, using features derived from the deterministic Weather Research and Forecasting (WRF) model (Skamarock et al., 2008); and (ii) the hourly forecast of sea surface temperature (SST) time series extracted from the operational Mercator global ocean analysis and forecast (Global ocean physics analysis and forecast). The results demonstrate a higher level of calibration compared to equivalent ensemble forecast techniques based on parametric forms for the CRPS.

The paper is organized as follows: Section 2 presents the used dataset and methodology, Section 3 shows the results, and Section 4 draws the conclusions. Finally, Section 5 reports data and software availability.

2. Methods

2.1. CRPS-based loss function

We implemented a CRPS-based loss function that, in a multi-output deep learning (DL) context, allows the final outputs of the neural network to be considered as an ensemble prediction. Specifically, the DL network output provides predictions $\hat{y} = \{\hat{y}_1\ \hat{y}_2\ \dots\ \hat{y}_M\}$, with each $\hat{y}_i$ representing an ensemble member and M the ensemble size. It is important to note that the input to the DL network is deterministic and not an ensemble, making the method fundamentally different from traditional ensemble prediction approaches.

2.1.1. Non-parametric approach

The CRPS is calculated here using the following non-parametric formula (Gneiting and Raftery, 2007):

$$\mathrm{CRPS} = \frac{1}{M}\sum_{i=1}^{M}\left|\hat{y}_i - y_{\mathrm{true}}\right| - \frac{1}{2M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}\left|\hat{y}_i - \hat{y}_j\right|. \tag{1}$$

In Eq. (1), the first term computes the average absolute error between each ensemble member $\hat{y}_i$ and the observed value y_{true}, promoting the accuracy of individual predictions. The second term calculates the average pairwise absolute differences among all ensemble members, encouraging diversity within the ensemble to accurately represent predictive uncertainty.

The optimization of the network involves adjusting the weights and biases θ of the network to minimize the CRPS loss function defined in Eq. (1). The logic of this optimization process is as follows. During training, the gradient of the CRPS loss with respect to θ can be estimated:

$$\nabla_\theta \mathrm{CRPS} = \frac{\partial \mathrm{CRPS}}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial \theta}.$$

The first term $\frac{\partial \text{CRPS}}{\partial \hat{y}}$ involves calculating the gradients of the loss with respect to the network outputs $\hat{y}$, while the second term $\frac{\partial \hat{y}}{\partial \theta}$ involves backpropagating these gradients through the network to update the parameters.

Using an optimizer, the network parameters θ can be updated in the direction that minimizes the CRPS loss (gradient descent):

$$\theta \quad \theta - \eta \nabla_\theta \text{CRPS}$$

where η is the learning rate. Though the gradient descent algorithm is particularly clear and well suited to explain the main principles underlying the optimization process, more advanced methods have been built on it in order to enhance the network's ability to find better solutions and converge more efficiently. Moreover, it is often the case where neural networks are not strictly differentiable and the concept of gradient itself has to be dealt with caution.

Overall, the optimization process guided by the CRPS loss function ensures that the network simultaneously learns to produce accurate predictions and to appropriately quantify uncertainty through the ensemble of outputs.

2.1.2. Parametric approach with Gaussian assumption

As a second strategy we consider as benchmark a neural network with exactly the same architecture, but now trained to minimize the closed-form expression that the CRPS follows under the strong assumption of Gaussian distribution of the prediction error (Gneiting and Raftery, 2007):

$$\text{CRPS}(\mu\ \sigma; y_{\text{true}}) = \sigma \left[\frac{1}{\sqrt{\pi}} - 2\phi(z) - z(2\Phi(z) - 1) \right]. \tag{2}$$

Here $z = \frac{y_{\text{true}} - \mu}{\sigma}$, and $\phi(\,)$ and $\Phi(\,)$ denote the probability density function (PDF) and cumulative distribution function (CDF) of the standard normal distribution, respectively. This loss function was already used in Rasp and Lerch (2018), even though our approach does not require an ensemble of predictions as input, but only relies on a deterministic forecast, which is one of the main novelties introduced in our approach.

In this parametric approach, the network optimization involves adjusting the weights θ to minimize the CRPS loss expressed in terms of μ and σ. These two parameters are computed from the predicted ensemble members $\hat{y} = \{\hat{y}_1\ \hat{y}_2\ \dots\ \hat{y}_M\}$ as their mean and standard deviation respectively, hence preserving the same output structure defined for the non-parametric case. Preserving the network architecture ensures a more fair comparison between methods. However, as proposed in Rasp and Lerch (2018), we also considered an alternative approach (afterward referred as *direct parametric*) where the output layer is defined by few nodes which are interpreted as parameters of the loss function (μ and σ in our cases). In this way, μ and σ are not computed from a generated ensemble, which can be obtained a posteriori by sampling from the predicted parametric distribution. In both cases, the neural network learns to adjust μ to minimize the difference between the predicted mean and the observed value y_{true}. Simultaneously, it adjusts σ to accurately represent the uncertainty in the predictions, as dictated by the CRPS loss.

A final remark has to be made here about using a parametric approach. Although any expression assumed a priori is necessarily a rough approximation, it is true that considering the CRPS formula for a distribution with a positive support, such as Gamma or Truncated Gaussian distributions, might be beneficial for semi-definite positive valued output variables, like the wind speed. However, we have explored also these possibilities experiencing technical issues related to slow convergence and loss divergence. Moreover, when we have managed to overcome these difficulties, for both Gamma and Truncated Gaussian distributions, the resulting model performs poorly. For this reason, we decide to not show the outcome of these attempts. This is an important aspect, which adds further value to the use of the non-parametric CRPS loss in terms of applicability and flexibility.

2.1.3. Comparison of the three strategies

The logic of the network optimization revolves around minimizing the CRPS loss to enhance the model's predictive performance and uncertainty quantification.

The non-parametric approach offers flexibility in modeling complex distributions without assuming a specific parametric form. The optimization process adjusts the ensemble outputs directly to minimize the CRPS, balancing accuracy and diversity among ensemble members. It could require a larger number of ensemble members to capture the full uncertainty, increasing computational demands.

The parametric and direct parametric approaches with Gaussian assumption simplify the problem by assuming a Gaussian distribution, which is completely identified by its mean and variance. Focusing the optimization on accurately estimating only this two numbers is possibly more efficient in computation, but may be less effective if the true distribution deviates significantly from normality.

In all cases the CRPS-based loss function plays a crucial role in guiding the optimization process, ensuring that the network not only predicts accurate mean values but also provides meaningful uncertainty estimates. The remainder of this paper is devoted to quantifying the potential pros and cons of these three strategies, ultimately drawing a definitive conclusion on their applicability.

2.2. Data collection

To demonstrate the potential of our model, we addressed two distinct tasks: the probabilistic prediction of the 10-meter wind speed and the hourly sea surface temperature (SST). For wind speed, we generate 24-hour ensemble forecasts using features extracted from the deterministic WRF model. For SST, we generate ensemble forecasts for the 24-hour time series based on a single input time series covering the preceding 24 h.

2.2.1. Wind data

The observed data used in this study are hourly surface-based wind data, at 10 m from the ground, from 39 different stations METARs (METeorologial Aerodrome Reports) scattered in the whole Italian territory (blue markers in Fig. 1) collected in the period 2020–2023. The METARs are a network of weather stations located mainly in airports and used for flight planning and weather forecasting (Ingleby, 2015).

To train our probabilistic neural network we used features coming from WRF model. It is a fully compressible non-hydrostatic, primitive-equation state-of-the-art numerical model with multiple nesting capabilities. A comprehensive description of the model formulation is given in Skamarock et al. (2008). The model configuration is as in Ferrari et al. (2021) with a resolution of 10 km covering the European region. The used features, reported in Table A.1, have been extracted by referring at the closest grid point to wind station, corresponding to timestamps of observed wind data. The train dataset has been made by concatenating all the stations data in the period 2020–2022, thus getting a dataset with a total number of about 700,000 samples; the test dataset is the remaining 2023 with a total number of about 250,000 samples.

2.2.2. SST data

To demonstrate our method in the alternative framework of time-series prediction, we randomly selected 22 geographical locations across the Mediterranean sea around the Italian territory (red markers Fig. 1). The selection was done seeking for diverse conditions, both in terms of sea basin and regularity of the SST signal. Data were harvested from the dataset (Global ocean physics analysis and forecast), which can be open accessed at the Copernicus Data Store. The global ocean analysis and forecast system provides hourly forecasts of surface fields, such as sea level height and sea surface temperature (SST), up to 10 days ahead on a horizontal grid with a regular spatial resolution of 1/12 degree in longitude and latitude. Since the time series are

Fig. 1. Map of Italy and its surrounding seas, showing the study's target locations. The 39 METAR stations on land are represented by blue markers, while the 22 marine locations used for sea surface temperature (SST) forecasting are marked in red.

aggregated over time to create a two-year sliding window, at the time of our analysis, we were able to collect data spanning from June 2022 to November 2024. For each location, we generated 21,339 24-hour time series by applying a one-hour rolling window to define the input and output. We adopt a standard approach for time-series forecasting (e.g., Lagomarsino-Oneto et al., 2023; Cavaiola et al., 2025), where our AI models are trained using samples organized to predict 24 h ahead, taking as input a preceding 24-hour time series. In this case, since observed SST data is not available, we use the SST predictions from the Copernicus Marine Service as the reference values. Comparisons are made between the 24-hour time series predicted by our AI models and the corresponding SST data from the Copernicus system. The training dataset consists of all 24-hour time series from the selected locations over the period 2022–2023, resulting in a total of 305,206 samples. The test dataset comprises the remaining time series from 2024, amounting to 164,252 samples.

2.3. Neural network architectures

To showcase the effectiveness of our method in generating well-calibrated probabilistic forecasts, we employed two straightforward neural network architectures tailored to the two forecasting tasks: hourly wind speed and hourly SST time series. Both architectures are multi-output neural networks, designed to ensure consistency in the number of trainable parameters, activation functions, regularizers, and initializers when transitioning from the Gaussian CRPS loss function to the non-parametric CRPS loss function. Note that, when using the direct parametric loss function, the neural network outputs directly μ and σ, introducing differences between architectures in number of trainable parameters, when shifting to this approach. Notably, the input to the networks is deterministic, not an ensemble, further distinguishing our approach from traditional ensemble-based methods.

For both forecasting tasks, we designed the architectures seeking for generalization across training and validation datasets. This was performed in a deterministic setup, where the Mean Squared Error (MSE) between target values and averages over the output nodes was used as loss function (this was done keeping the number of output nodes $M = 100$, which is the highest value that we tested). Afterward, we performed a fine-tuning of those hyperparameters that are directly involved in the optimization process. Specifically, the best configuration is sought among the learning rate (selecting it in a range of 0.001 to 0.00001), optimizer (selecting one from Adam, RMSprop or SGD) and batch size (selecting one from 256, 512 and 1024). For both tasks and for both parametric and non-parametric loss functions we selected as best configurations, the one yielding the minimum product of the loss function (CRPS) and the Δ index (see Section 3.3), evaluated on the validation set. The same approach of fine-tuning is performed to the neural networks using the direct-parametric loss function, evaluating the metrics between observed values and predicted mean prediction, on the validation sets.

2.3.1. Neural network architecture for ensemble wind forecasting

Figs. 2 provides a schematic representation of the neural network used for ensemble wind speed forecasting. The network consists of three hidden layers and a multi-output layer, with the number of nodes and activation functions detailed in the figure. The CRPS loss function, applicable in both parametric and non-parametric cases, is applied to the final layer. The nodes in the multi-output layer can be interpreted as members of the ensemble forecast. To evaluate the impact of ensemble size on probabilistic forecasting skill, we varied the number of output nodes. In this case, the fine-tuning of optimization parameters selected the Adam optimizer, with an initial learning rate of 0.0002, for both the parametric and the non-parametric loss, while for the batch size, 512 is used for the former and 1024 for the latter. For

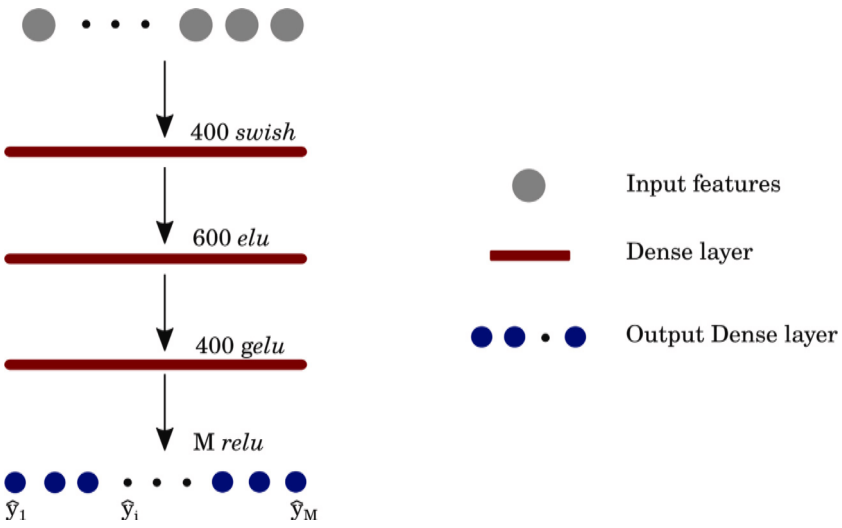


Fig. 2. Architecture of the neural network used for wind forecasting. The legend specifies the role of each layer. The numbers above each bar represent the number of nodes in the output, while the activation function used in each layer is also indicated.

both loss functions, the training employs early stopping with a patience of 10 epochs and learning rate reduction, with a patience of 5 epochs, applying a reduction factor of 0.5. The validation loss threshold for triggering learning rate reduction and early stopping is set to 0.0001. As a pre-processing, all input features are normalized between 0 and 1 before training.

2.3.2. Neural network architecture for ensemble SST forecasting

For the time series forecasting task, we used the architecture illustrated in Fig. 3. The input sequence first passes through a one-dimensional convolutional layer with 200 filters and symmetric padding. This is followed by a bidirectional LSTM layer consisting of 200 LSTM cells, which is stacked on top of another one-dimensional convolutional layer with 150 filters and symmetric padding. The final output layer is a one-dimensional convolutional layer with symmetric padding and M filters. Similar to the wind forecasting case, we vary the value of M, and the CRPS loss function is applied to this final layer. The last multi-output layer can be interpreted as members of an ensemble forecast. Before feeding into the neural network all data were normalized between 0 and 1. The fine-tuning process selected the Adam algorithm with a starting learning rate of 0.001 and batch size of 256, for both parametric and non-parametric loss functions. The networks are trained using early stopping with a patience of 10 epochs and learning rate reduction with a patience of 5 epochs, employing a reducing factor of 0.5. The validation loss threshold for learning rate reduction and early stopping is set at 0.0001.

2.3.3. Neural networks for direct parametric loss function

To test the performance of the direct parametric loss function, we used the same architectures reported in Sections 2.3.1 and 2.3.2, for both wind and SST forecasting, respectively. The only difference is in the number of outputs, where in this case they are set to 2, μ and σ respectively. We use the same pre-processing of the data described above, and for both forecasting tasks we performed fine tuning of hyperparameter such as learning rate, optimizer and batch size. The fine tuning process selected Adam optimizer with learning rate of 0.0001 for wind forecast, and Adam optimizer with learning rate of 0.0002 for SST forecast, both with a batch size of 512. With this set up, to have a consistent comparisons between the methods we show, the ensemble members are obtained by sub-sampling M values form the Gaussian distribution with μ and σ output from the neural network.

3. Results

In this section, we present the results of the probabilistic forecasts for the 24-hour-ahead wind speed and SST predictions. It is worth emphasizing that the main goal of this study is to evaluate the quality of the full probabilistic ensemble generated by our method, rather to assess the accuracy of deterministic forecasts, such as the ensemble mean. The forecast evaluation has been performed using a combination of statistical indices widely employed in the scientific literature, highlighting the importance of complementary metrics to provide a comprehensive assessment of forecast skill. Specifically, the probabilistic forecast performance will be evaluated using the Continuous Ranked Probability Score (CRPS), the Δ index, and the binary cross-entropy. These metrics capture different aspects of the ensemble forecast's quality, such as calibration, and the ability to assign accurate probabilities to threshold-based events.

In addition to the probabilistic assessment, the deterministic skills of the forecasts will be analyzed using the Normalized Root Mean Square Error (NRMSE) and the Pearson correlation coefficient, calculated for each location during the test period. While these deterministic metrics provide valuable insights into the accuracy of the ensemble mean, they are insufficient to capture the full probabilistic nature of the forecasts. Thus, the inclusion of complementary probabilistic metrics ensures a more complete and robust evaluation of the forecast's overall performance.

3.1. Assessing the overall probabilistic skill of the ensemble prediction system

The overall probabilistic skills of our method and related benchmarks have been firstly quantified by calculating the CRPS on test data. For both tasks, we analyze the variation of the CRPS with the ensemble size (Fig. 4), defined by the number of output nodes in the model (or by the number of resampled values for the direct parametric method). We also associate 95 confidence intervals to the estimated CRPS values by performing a 10^4 bootstrap resampling of the location-averaged CRPS values. For wind forecasts (Fig. 4, panel a), the non-parametric approach consistently outperforms the parametric one across all ensemble sizes, as indicated by lower CRPS values. Instead, slight but still significative differences emerge between non-parametric and direct parametric methods. In the SST case (Fig. 4, panel b), the performance of the non-parametric method remains the best, but the parametric provides similar performances. Differently from the wind

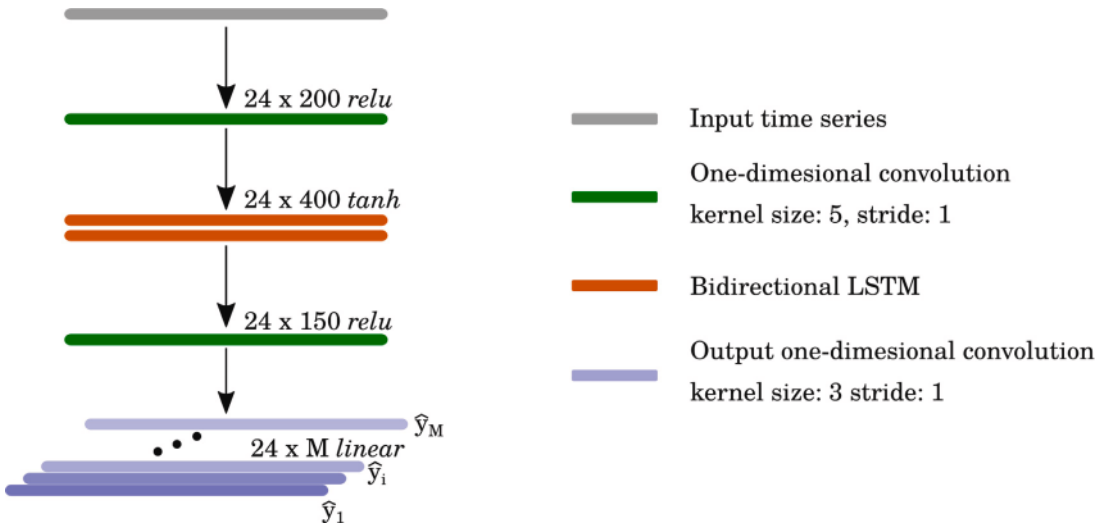


Fig. 3. Architecture used for 24-hour time series SST forecast. The legend specifies the role of each layer. The numbers above each bar represent the output shape of each layer in the format $24 \times M$, where 24 is the time series length and M represents the latent features. The activation function used in each layer is also indicated.

case, the direct parametric is now significantly worse. However, in terms of accuracy and overall probabilistic structure of the forecast, our approach turns out to be optimal in all cases, further confirming its high flexibility. Lastly, it is interesting to notice that, in terms of ensemble size and from a CRPS perspective, the parametric is more unstable than our method.

3.2. Forecast reliability

Forecast reliability, closely related to calibration, assesses the agreement between predicted probabilities and observed relative frequencies (Silva Filho et al., 2023; Bothwell, 2010). For the tasks of the present study, a forecast is perfectly calibrated if, for cases predicted to have, say, x probability of exceeding a given wind speed or SST threshold, the observed frequency of such events within those cases is indeed x .

Following this idea, we have used percentiles from the observed distribution of our target variable as thresholds to define classification problems, where classes 1 and 0 correspond to observing higher and lower values respectively. This procedure allows to construct reliability diagrams for these surrogate tasks. Indeed, by evaluating the proportion of ensemble members exceeding the threshold, each of our ensemble of M predictions can be immediately translated into a probability for class 1.

These probabilities have been partitioned into 10 disjoint bins between 0 and 1. For each bin, the observed frequency of class 1 events has been calculated and plotted against the mean predicted probability for that bin. The diagonal reference line indicates perfect reliability, while deviations reflect calibration issues: points below the diagonal indicate over-prediction, and points above indicate under-prediction. This diagram is shown in Fig. 5 and provides a visual evaluation of the model's reliability for the defined event. Panels (a) and (b) correspond to the wind and SST forecasts, respectively. To ensure a comprehensive evaluation of the model's performance, we have selected four threshold values corresponding to the 25th, 50th, 75th, and 95th percentiles. These thresholds capture a range of conditions from lower to extreme values, allowing us to assess the model's reliability across different regimes. Reliability diagrams (Fig. 5) draw a picture that is coherent with the behavior showed by the CRPS. In the case of wind forecast (Fig. 5, panel a), the direct parametric and the non-parametric methods perform better than the parametric. Remarkably, the non-parametric case is superior in capturing extreme values (25th and 95th percentiles). In the SST case (Fig. 5, panel b), both the parametric and non-parametric methods behave similarly, with curves very close to

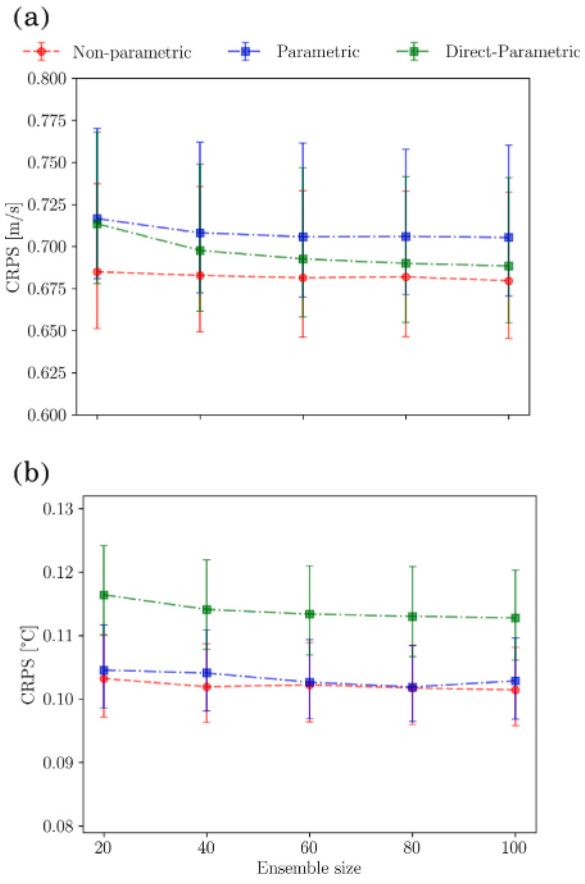


Fig. 4. CRPS evaluated on test data as function of the ensemble size. Colors indicate different strategies: non-parametric CRPS loss function (red), parametric (blue) and direct parametric (green). Error bars represent 95 confidence intervals obtained via bootstrap resampling. Different scenarios emerge for the wind (panel a) and the SST (panel b) forecasts. In both cases the non-parametric method turns out to be an optimal and more stable strategy.

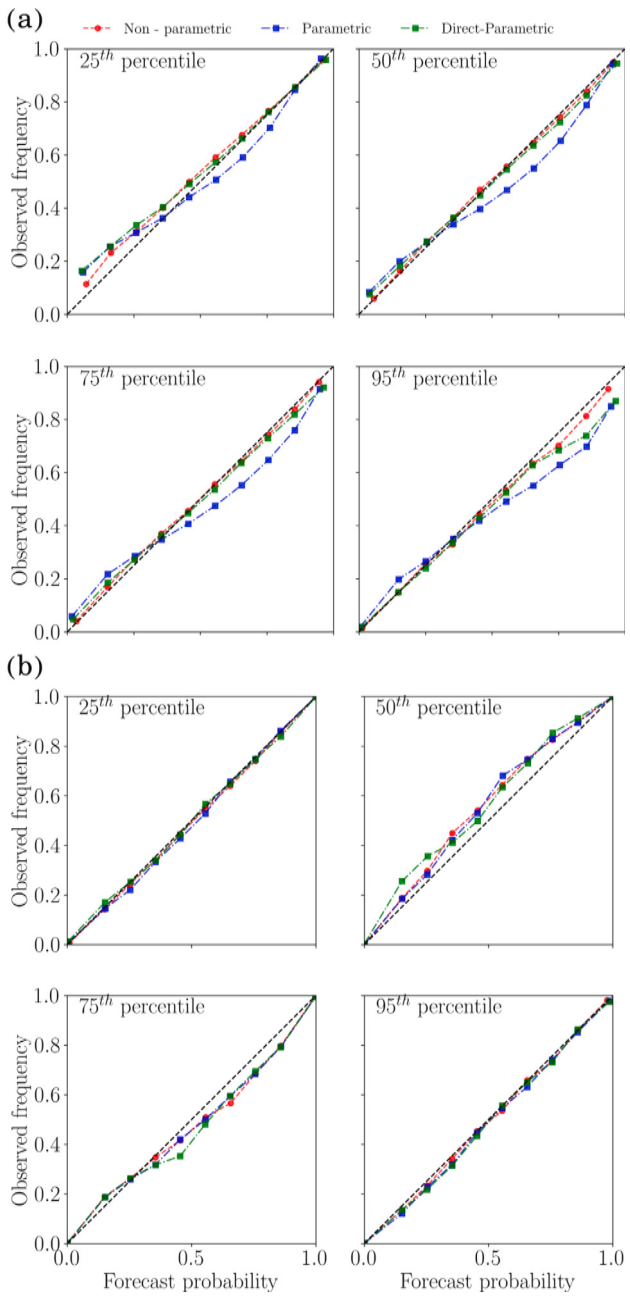


Fig. 5. Panel (a): reliability diagrams for the network with 100 output nodes applied to wind forecasts. The cases correspond to four different wind thresholds defined by the 25th, 50th, 75th, and 95th percentiles (from the ground truth marginal distribution of target values). The results are shown for the non-parametric CRPS-loss function (red line), the parametric CRPS-loss function (blue line) and the direct parametric CRPS-loss function (green line). The diagonal represents the optimal reliability. The x-axis indicates the predicted probability, while the y-axis shows the observed frequency. Panel (b): as in panel (a) but for the 24-hour SST forecasts.

the diagonal, indicating a high degree of reliability in all considered scenarios.

3.3. More on forecast calibration: Rank histogram

In the context of ensemble forecasts, calibration is often evaluated using the Rank Histogram (Hamill, 2001), which assesses whether the ensemble members $\{\hat{y}_1 \ \hat{y}_2 \ \cdots \ \hat{y}_M\}$ accurately represent the uncertainty of the forecast. Specifically, the rank of the observed value y within the sorted ensemble members is calculated for each forecast instance. For an ensemble of M members, let:

$$\hat{y}_{(1)} \quad \hat{y}_{(2)} \quad \cdots \quad \hat{y}_{(M)}$$

denote the sorted ensemble members. The rank of the observation y is defined as:

$$\text{Rank}(y) = 1 + \sum_{i=1}^{M} (\hat{y}_{(i)} \quad y)$$

where ( ) is the indicator function, equal to 1 if the condition inside is true, and 0 otherwise. The rank Rank(y) lies between 1 and $M+1$, where a rank of k indicates that the observation falls between the $(k-1)$-th and kth ensemble members, or below the first member if $k = 1$, or above the last member if $k = M + 1$.

The Rank Histogram is constructed by counting the frequencies of the ranks Rank(y) across all forecast instances and visualizing these counts as a histogram. If the ensemble is well-calibrated, the ranks should be uniformly distributed across [1 M + 1], resulting in a flat histogram. Deviations from uniformity reveal systematic biases or misrepresentation of uncertainty. A U-shaped histogram indicates that the ensemble is under-dispersed (too narrow), meaning the observation frequently falls outside the range spanned by the ensemble. A hill-shaped histogram suggests the ensemble is over-dispersed (too wide), with the observation clustering near the center of the ensemble range. An asymmetric histogram indicates a bias, where the ensemble systematically under- or overestimates the observed values.

To quantify deviations from uniformity, we use the reliability index (Delle Monache et al., 2006; Casciaro et al., 2022), here denoted by Δ, which measures the deviation of the Rank Histogram from a uniform distribution:

$$\Delta = \sum_{k=1}^{N_{bin}} \left| \frac{n_k}{N} - \frac{1}{N_{bin}} \right| \tag{3}$$

where n_k is the count of ranks in the kth bin of the histogram, N is the total number of forecasts, and $\frac{1}{N_{bin}}$ is the expected frequency for each bin under uniformity. A lower Δ value indicates better calibration, as it reflects a histogram closer to uniformity.

Fig. 6 shows the rank histogram for the wind speed forecast at three example locations. The selection has been done considering Δ values from all locations and considering the 25th, 50th, and 75th percentiles of the corresponding distribution. Fig. 7 analogously illustrates examples for the SST forecast. The rank histograms presented for wind speed and SST forecasts reveal crucial insights into the calibration of the parametric and non-parametric CRPS loss functions. Starting with the wind speed forecasts, it is evident that the non-parametric loss achieves superior calibration. The histograms for the 25th, 50th, and 75th percentiles show distributions that are closer to uniformity, with Δ indices of 0.05, 0.07, and 0.08, respectively. These low Δ values indicate minimal deviation from the ideal flat histogram, suggesting that the non-parametric loss captures the forecast uncertainty effectively. While there are slight deviations, such as minor peaks at the edges for some percentiles, these issues are relatively insignificant, and the overall calibration is strong.

In contrast, the parametric loss for wind speed demonstrates substantial under-dispersion, as highlighted by the pronounced U-shaped patterns in the histograms. The Δ indices for the parametric loss, ranging from 0.33 to 0.46, are considerably higher than those for the non-parametric loss, signaling poor calibration. This under-dispersion implies that the ensemble members fail to span the observed values adequately, causing the observations to frequently fall outside the range of the ensemble. The issue becomes particularly severe at the 75th percentile, where the inability of the parametric loss to capture extreme events is evident. However, the direct-parametric approach shows a better calibration even though, its Δ indices are still higher than the non-parametric. Particularly, for the selected locations in Fig. 6, the

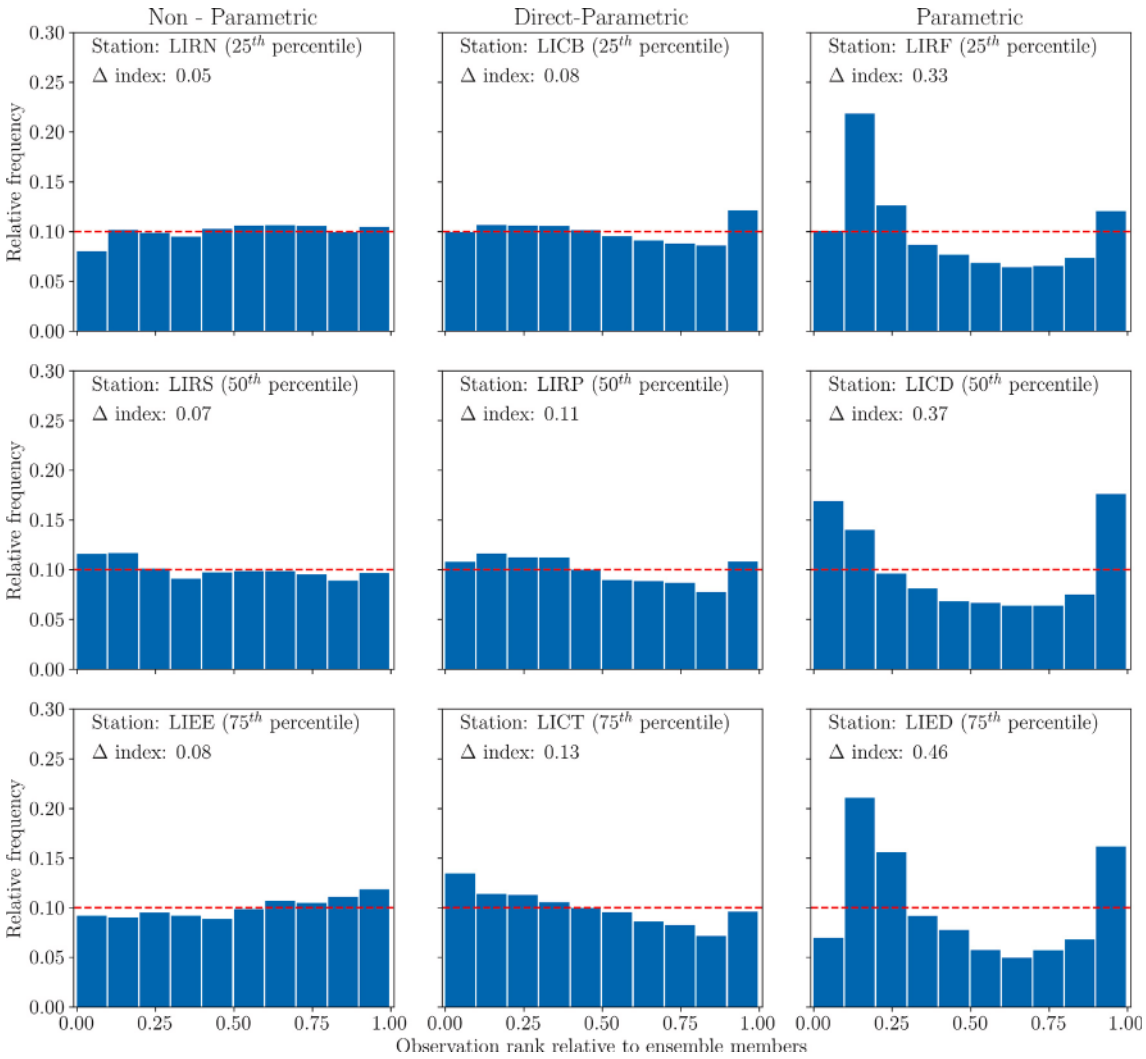


Fig. 6. Rank histograms for the network with 100 output nodes applied to wind forecasts. Each column of panels refers to one of the considered approaches (non-parametric, direct-parametric and parametric). The diagrams correspond to three stations selected based on the 25th percentile (upper panel), 50th percentile (middle panel), and 75th percentile (lower panel) of the Δ index. The x-axis represents the PIT values divided into 10 bins, and the y-axis shows the relative frequency of predictions in each bin. The red line at $y = 1/N_{bin}$ indicates perfect uniformity, representing the expected frequency for an ideal probabilistic forecast. The label 'Station' denotes the METAR code of the selected location.

direct-parametric method has a tendency to overestimate high wind speeds while underestimating lowest values.

Turning to the SST forecasts, the non-parametric loss remains the best choice, but the difference with the parametric method is now less pronounced. Differently from the wind case, the direct-parametric approach is now worse. Moreover, in the selected examples, both parametric and direct parametric methods show a U-shape, leaning to underestimate output distributions symmetrically. The rank histograms for the non-parametric loss remain relatively uniform, with Δ indices of 0.08, 0.10, and 0.12 for the 25th, 50th, and 75th percentiles, respectively. These values, while low, are slightly higher than those observed for wind speed, indicating that the non-parametric loss struggles more with SST forecasts. Mild over-dispersion is evident in some histograms, particularly at the 50th and 75th percentiles, where hill-shaped patterns suggest that the ensemble members might overestimate the forecast uncertainty. Nevertheless, the non-parametric loss provides a consistent and robust representation of SST variability, despite these minor issues.

The parametric loss for SST, as with wind speed, shows significant calibration problems. The Δ indices range from 0.1 to 0.17, higher than those for the non-parametric approach. The observed deviations from uniformity in the histograms highlight the limitations of the parametric approach in capturing the range and spread of SST uncertainty.

In Fig. 8, we report the behavior of the average, across all locations, of the Δ index as a function of the ensemble size for wind forecasts (panel a) and SST forecasts (panel b). For wind forecasts, the Δ index highlights systematic differences across methods. The Δ values for the parametric approach remain significantly higher, indicating poorer calibration of the ensemble. In contrast, the non-parametric approach exhibits small values of Δ across different ensemble sizes, demonstrating its robustness and superior calibration. The direct-parametric approach shows calibration skills closer to the non-parametric, but still systematically lower. Δ index values of SST forecast also show a systematic hierarchy of methods, with the non-parametric loss still performing best. Noticeably, this behavior is stable to variations in the ensemble size.

3.4. Measuring the alignment between predicted probabilities and observed outcomes

While reliability metrics and rank histograms provide valuable insights into calibration and the probabilistic spread of the ensemble, they do not directly evaluate the model's ability to assign accurate

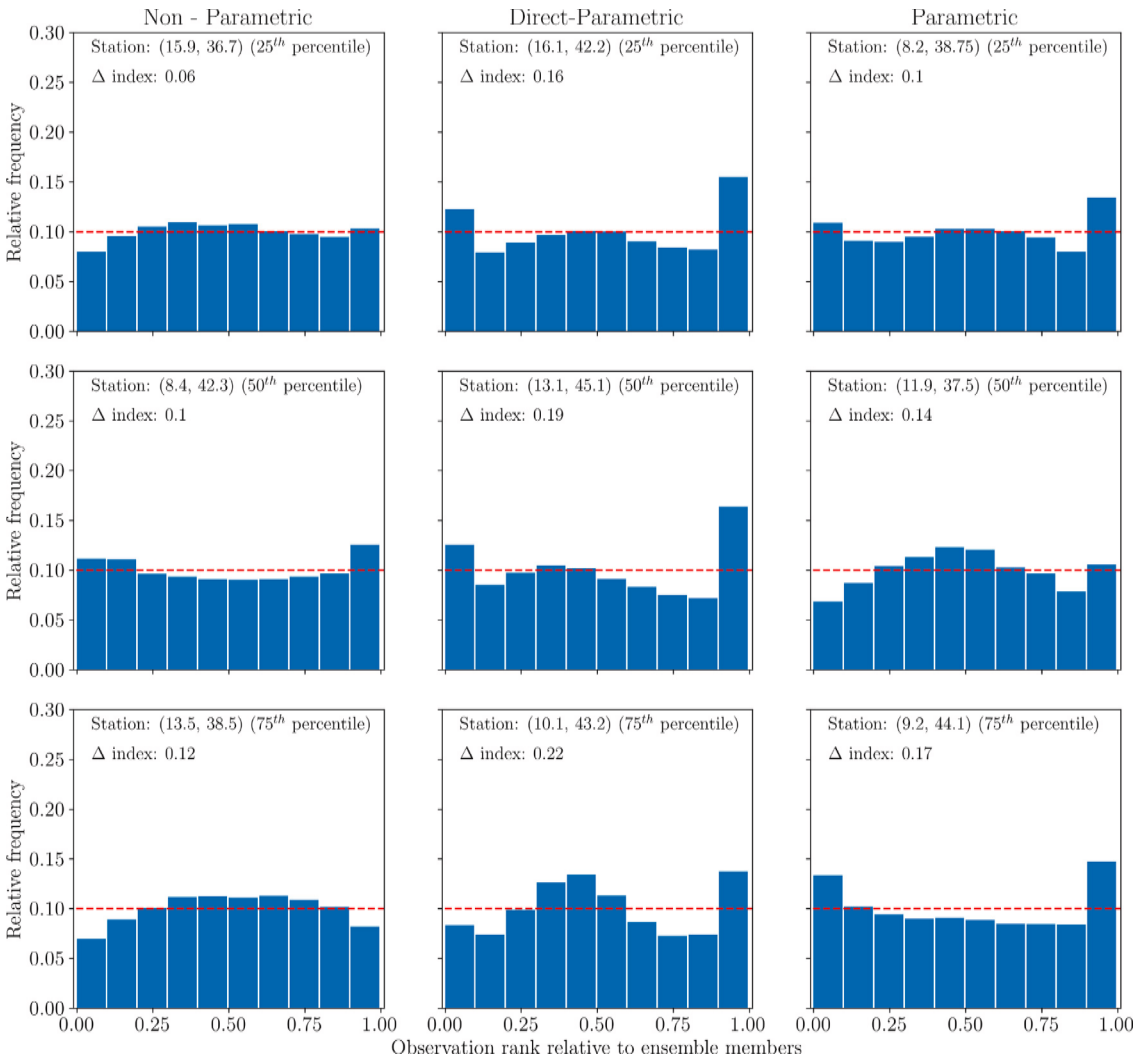


Fig. 7. Rank histograms for the network with 100 output nodes applied to 24-hour SST forecasts. Each column of panels refers to one of the considered approaches (non-parametric, direct-parametric and parametric). The diagrams correspond to three locations that were selected based on the 25th percentile (upper panel), 50th percentile (middle panel), and 75th percentile (lower panel) of the Δ index. The x-axis represents the PIT values divided into 10 bins, and the y-axis shows the relative frequency of predictions in each bin. The red line at $y = 1/N_{bin}$ indicates perfect uniformity, representing the expected frequency for an ideal probabilistic forecast. Each panel reports the longitude-latitude coordinate of the selected location.

probabilities to binary events, such as exceeding a given threshold. Cross-entropy addresses this gap penalizing forecasts that assign low probabilities to events that actually occur (underconfidence) or assign high probabilities to events that do not occur (overconfidence). This makes it a critical tool for assessing how well the probabilistic forecasts capture class-specific behavior, especially in applications like extreme event prediction.

In the panel (a) of Fig. 9, which focuses on wind forecasts, we observe that the non-parametric CRPS loss consistently achieves similar or lower cross-entropy values across all ensemble sizes and thresholds (25th to 95th percentiles) compared to the direct-parametric approach. Instead, the parametric approach is slightly better at accurately representing the probability of exceeding the thresholds of 25th and 50th percentiles of wind speed. However, the non-parametric loss is again significantly better than the parametric for higher percentiles (75th and 95th), where accurate probabilistic representation is crucial for extreme event prediction.

In the panel (b) of Fig. 9, which pertains to SST forecasts, the cross-entropy values are lower overall compared to the wind forecasts, reflecting the generally smoother and less variable nature of SST predictions. In this case, the non-parametric approach systematically outperforms both parametric and direct-parametric, though differences are less pronounced than in the wind forecast case. This indicates that all methods perform reasonably well in predicting SST exceedance probabilities, but the non-parametric approach still maintains a slight advantage.

3.5. Assessing the skills of the mean ensemble prediction

The performance of the mean ensemble prediction, as assessed through the normalized root mean square error (NRMSE) and the Pearson correlation coefficient, reveals no significant difference between the two parametric and the non-parametric approaches. Specifically, for wind forecasts, all models achieve an NRMSE of about 37 . The Pearson coefficient is 69 for the direct-parametric approach and about

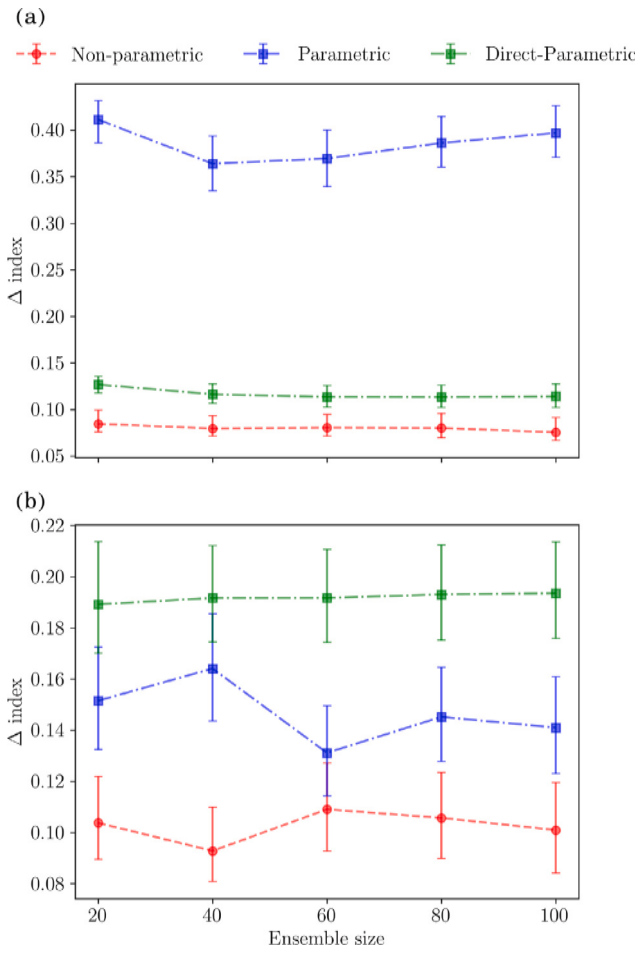


Fig. 8. Δ index for wind forecasts (upper panel) as a function of the number of output nodes in the model. The red line corresponds to the non-parametric CRPS-loss function, the blue line represents the parametric CRPS-loss function while the green line shows the direct parametric CRPS-loss function. Error bars indicate 95 confidence intervals obtained via bootstrap resampling. Panel (b): as in panel (a) but for the SST forecasts.

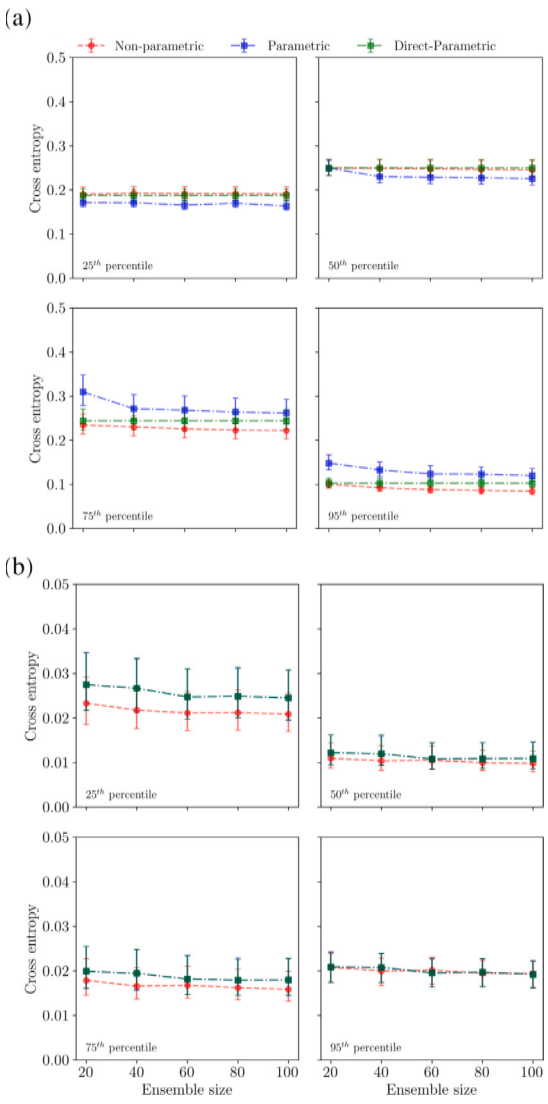


Fig. 9. Panel (a): cross-entropy of class 1 as a function of the number of ensemble members for 24-hour (hourly) wind forecasts. Class 1 corresponds to exceeding the wind threshold defined by the respective percentiles (25th, 50th, 75th, and 95th), which are calculated from the distribution of observed wind speeds. Results are shown for the non-parametric CRPS-loss function (red curve), the parametric CRPS-loss function (blue curve) and the direct parametric CRPS-loss function (green curve), where the optimal cross-entropy value, theoretically, is zero. Error bars indicate 95 confidence intervals obtained via bootstrap resampling. Panel (b): as in panel (a) but for the 24-hour SST forecasts.

70 for the other two (reported values are averages across all locations). Similarly, for SST forecasts, all models yield identical metrics, with an NRMSE of about 82 and a Pearson coefficient of about 70 . These results suggest that, if one were to base their evaluation solely on the skill of the ensemble mean, there would be no discernible advantage in choosing one model over the other. However, such a conclusion would be incomplete and misleading. As demonstrated by the probabilistic metrics discussed earlier, including the CRPS, Δ index, rank histograms, and cross-entropy, the non-parametric loss consistently outperforms parametric models in terms of calibration and overall probabilistic skill. Thus, while the assessment of the ensemble mean provides valuable information about central tendency, it fails to capture the probabilistic quality of the forecast, which is crucial for applications requiring a comprehensive representation of uncertainty. This highlights the importance of incorporating a broader range of metrics when evaluating ensemble forecast performance.

4. Conclusions

In this study, we have introduced a novel deep learning framework, CRPS-Net, designed to generate ensemble forecasts directly from deterministic inputs and through only a single Deep Neural Network. By employing the Continuous Ranked Probability Score (CRPS) as a loss function in its non-parametric formulations, the proposed method has addressed some important limitations of traditional ensemble approaches, significantly reducing computational costs while maintaining robust probabilistic forecast skill. Our results have demonstrated that the non-parametric CRPS loss consistently outperforms parametric approaches, particularly for wind forecasts, where calibration and probabilistic reliability are critical. The reliability diagrams and rank histograms analyses have shown that the non-parametric loss achieves superior calibration, effectively capturing the forecast uncertainty without relying on strong assumptions about the error distribution. Quantitative metrics, including the Δ index, have further highlighted the robustness of the non-parametric strategy across different ensemble sizes and thresholds.

We have validated the versatility of CRPS-Net through two distinct applications: 24-hour wind forecasts using deterministic WRF model

outputs and hourly time-series forecasts of sea surface temperature (SST) from ocean forecast datasets. In both cases, the framework has shown consistent performance, offering a scalable and reliable alternative to traditional ensemble prediction systems, particularly in operational settings where the computational burden of simulation ensembles is often unsustainable.

Looking ahead, the non-parametric CRPS loss function presents significant potential for application in even more complex architectures, such as image-based convolutional neural networks. This opens the door to extending the methodology to spatially and temporally rich datasets, enabling the estimation of uncertainty in high-resolution forecasts for a variety of geophysical fields. By integrating this strategy into advanced deep learning architectures, it could provide critical tools for uncertainty quantification in applications ranging from atmospheric science to oceanography and beyond. Future research should also explore ways to further optimize the neural network architectures and evaluate the approach in diverse forecasting domains. In the specific context of hydrology, a particularly promising synergy lies in combining CRPS-Net's ability to generate probabilistic forecasts with AutoGluon's ensemble learning capabilities (Tu et al., 2024) for optimizing the selection of meteorological inputs and calibrating models to better capture hydrological processes. In way of example, while AutoGluon excels at identifying the most relevant meteorological drivers and combining diverse machine learning models, CRPS-Net could provide probabilistic outputs that enhance uncertainty quantification and improve decision-making in hydrological applications. Together, these methods could address the challenge of reconstructing and forecasting natural flows by ensuring robust calibration across varying conditions.

CRPS-Net could also complement existing multi-model approaches, such as those described by Simmons and Splinter (2022) which leverage weighted ensembles to enhance predictions for extreme coastal erosion events. Specifically, CRPS-Net's ability to adaptively model error distributions without predefined assumptions would address potential biases in the ensemble's individual components, thereby refining predictions for both high-impact and moderate events.

5. Software and data availability

Name of software: CRPS-Net
Developers: Mattia Cavaiola and Daniele Lagomarsino-Oneto
Contact: mattia.cavaiola@gmail.com and mattia.cavaiola@cnr.it
Date first available: December 21, 2024
Program language: Python
Software required: all python libraries can be found at https://github.com/Mattiads1/CRPS-Net/blob/main/requirements.txt. All those libraries can be accessed freely through the cloud.
Loss function source code at: https://github.com/Mattiads1/CRPS-Net/blob/main/CRPSLoss.py
Documentation: Documentation for application, installation, testing, and deployment can be found at https://github.com/Mattiads1/CRPS-Net/blob/main/README.md
SST data used in this work can be freely accessed from the cloud (see Global ocean physics analysis and forecast). WRF data and METAR wind observations will be made available upon reasonable request to the authors.

CRediT authorship contribution statement

Mattia Cavaiola: Writing – review & editing, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Daniele Lagomarsino Oneto:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Data curation. **Andrea Mazzino:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Formal analysis.

Table A.1
Features used for training the neural network applied to hourly wind speed forecasts. The features correspond to values extracted from the WRF model at the nearest grid point to the observation stations.

List of WRF features used to train the AI networks
$u10$ - u component wind velocity at 10 m
$v10$ - v component wind velocity at 10 m
$u30$ - u component wind velocity at 30 m
$v30$ - v component wind velocity at 30 m
$u50$ - u component wind velocity at 50 m
$v50$ - v component wind velocity at 50 m
$u80$ - u component wind velocity at 80 m
$v80$ - v component wind velocity at 80 m
$u100$ - u component wind velocity at 100 m
$v100$ - v component wind velocity at 100 m
$u305$ - u component wind velocity at 305 m
$v305$ - v component wind velocity at 305 m
wg - wind gust
$T2m$ - air temperature at 2 m
$T10$ - air temperature at 10 m
$T30$ - air temperature at 30 m
$T50$ - air temperature at 50 m
$T80$ - air temperature at 80 m
$T100$ - air temperature at 100 m
$T305$ - air temperature at 305 m
blh - boundary layer height
p - surface pressure
sw - incoming short wave
lw - incoming long wave
$ws10$ - wind velocity at 10 m
$dir10$ - wind direction at 10 m
$ws30$ - wind velocity at 30 m
$dir30$ - wind direction at 30 m
$ws50$ - wind velocity at 50 m
$dir50$ - wind direction at 50 m
$ws80$ - wind velocity at 80 m
$dir80$ - wind direction 80 m
$ws100$ - wind velocity at 100 m
$dir100$ - wind direction at 100 m
$ws305$ - wind velocity at 305 m
$dir305$ - wind direction 305 m
$\cos(2\pi h/24)$ - h being the hour of the day
$\sin(2\pi h/24)$ - h being the hour of the day
lat - latitude
lon - longitude

Declaration of competing interest

The authors have declared no conflict of interest.

Acknowledgments

Compagnia di San Paolo is gratefully acknowledged for its financial support (project AIxtreme, under the "Intelligenza Artificiale" 2022 call) for the computational infrastructure where the calculations were performed. We also extend our gratitude to the PNRR project RAISE - "Robotics and AI for Socio-economic Empowerment" for funding the research activities, specifically supporting human resources. D.L. was supported by EU—Next Generation EU Mission 4 "Education and Research" Project IR0000032—ITINERIS—Italian Integrated Environmental Research Infrastructures System—CUP B53C22002150006. We also acknowledge the valuable discussions with Francesco Ferrari.

Appendix A. Features used for training the neural network for wind forecasting

Table A.1 reports the features used to train the neural network for wind speed forecasting. The WRF model outputs used in this study are from the operational setup run at DICCA - UNIGE (www.meteocean.science). The model configuration employed, here with a horizontal resolution of 10 km, has been utilized in numerous studies focusing on

the Mediterranean region (see, e.g., Cassola et al., 2015, 2016; Ferrari et al., 2021 among the others).

Appendix B. Indices to assess the skill of ensemble mean predictions

The error indices used here are the normalized root mean square error (RMSE) and the correlation coefficient. Considering Y_n the nth value of the ground truth and X_n the corresponding nth forecast (here corresponding to the mean of the M ensemble members), the NRMSE is defined as:

$$NRMSE = \sqrt{\frac{\sum_{n=1}^{N}\left(X_n - Y_n\right)^2}{\sum_{n=1}^{N} Y_n^2}} \tag{B.1}$$

where N is the number of ground truth-forecast pairs over which the performance is evaluated. In the case of wind forecast N is simply the test set size of each station. Since for SST predictions we are providing a multi-horizon forecast from 1 to 24 h ahead, all deterministic metrics are computed on each predicted window, hence N=24 in this case, and then averaged over all test samples.

The correlation coefficient, C, is a measure of linear dependence between two variables (Wilks, 2011) and takes values between −1 and 1, where 1 represents the maximum correlation, −1 the maximum anti-correlation. In plain formula (Lee Rodgers and Nicewander, 1988),

$$C = \frac{\sum_{n=1}^{N}(X_n - \overline{X})(Y_n - \overline{Y})}{N\sigma_X\sigma_Y} \tag{B.2}$$

with:

$$\sigma_X = \sqrt{\frac{\sum_{n=1}^{N}(X_n - \overline{X})^2}{N}} \tag{B.3}$$

$$\sigma_Y = \sqrt{\frac{\sum_{n=1}^{N}(Y_n - \overline{Y})^2}{N}} \tag{B.4}$$

where $\overline{X}$ and $\overline{Y}$ are the mean values of X and Y.

Appendix C. Computational efficiency

Fig. C.10 shows the computation time in seconds (sec) required for training the AI model for wind speed (top panel) and SST (bottom panel) prediction, and for the two implementations of the Parametric and Non-Parametric loss functions. The computational tests are performed on an HPC system equipped with one GPU NVIDIA-H100 of 96 Gb of memory and 96 CPUs AMD EPYC 9224. Performance experiments are performed by running 10 train processes for each ensemble size, to account for uncertainty (error bars in Fig. C.10). Each train has been run for 10 epochs and batch size of 1024.

We have made this analysis only for the parametric and non-parametric approaches. Indeed, the direct-parametric method involves a slightly different architecture, which makes unfeasible to isolate the gain that is actually due to a different choice of the loss function. In the case of wind forecast, the computational time needed to train the neural network is systematically lower, while for the SST case a little advantage emerges as the ensemble size grows. However, the purpose of this analysis is to provide an idea of resources required for the proposed approach (and its variations). Indeed, once we have found out that the strategy produces accurate and well calibrated outputs, the time required to train the model is of the order of a few tens of seconds per 10 epochs (for the chosen batch size). This computational burden is not even comparable with the amount of resources needed to produce an ensemble of simulations. Furthermore, in operative forecasts, such simulated ensembles need to be continuously run. On the contrary, after the training stage, methods which are based on a single running simulation with neural-network-based generation of the ensemble have only the additional burden of using the network for inference.

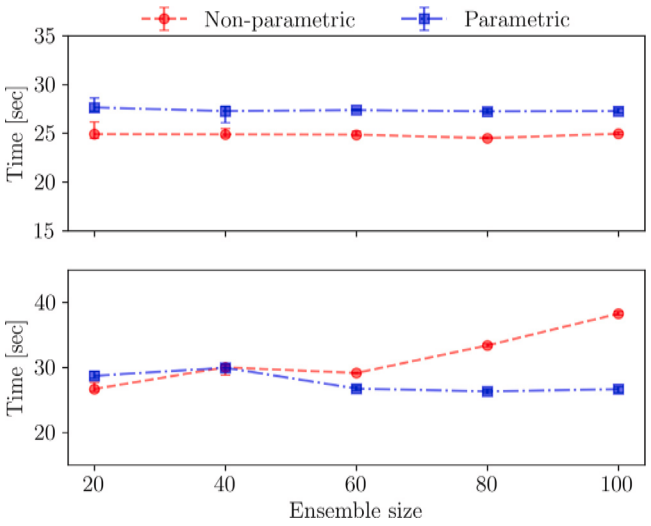


Fig. C.10. Execution time (sec) required for training the AI models for wind speed forecasting (top panel) and SST prediction (bottom panel), as a function of ensemble size, for both Non-Parametric (red line) and Parametric (blue line) loss implementation. Error bars indicate 95 confidence intervals obtained over 10 train process of 10 epochs via bootstrap resampling.

Data availability

Data will be made available on request.

References

Balmaseda, M.A., Bell, M.J., Ritchie, H., 2015. The role of the ocean in climate predictability. J. Clim. 28 (3), 891–892.

Bothwell, P., 2010. Evolution of the experimental/automated perfect prog lightning forecasts at the storm prediction center. In: Third International Lightning Detection Conference, April 21-22. Orlando, FL, Vaisala, Inc., Tucson.

Buizza, R., et al., 1999. The potential value of ensemble prediction as a tool for forecasting. Q. J. R. Meteorol. Soc. 125 (560), 2887–2908.

Casciaro, G., Ferrari, F., Cavaiola, M., Mazzino, A., 2022. Novel strategies of ensemble model output statistics (EMOS) for calibrating wind speed/power forecasts. Energy Convers. Manage. 271, 116297. http://dx.doi.org/10.1016/j.enconman.2022.116297, https://www.sciencedirect.com/science/article/pii/S0196890422010755.

Cassola, F., Ferrari, F., Mazzino, A., 2015. Numerical simulations of mediterranean heavy precipitation events with the wrf model: A verification exercise using different approaches. Atmospheric Res. 164- 165, 210–225, https://www.sciencedirect.com/science/article/pii/S0169809515001696.

Cassola, F., Ferrari, F., Mazzino, A., Miglietta, M.M., 2016. The role of the sea on the flash floods events over liguria (northwestern italy). Geophys. Res. Lett. 43 (7), 3534–3542. http://dx.doi.org/10.1002/2016GL068265.

Cavaiola, M., Marini, Simone, Magaldi, M.G., Mazzino, A., 2025. Ai-driven 6-hour ahead nowcasting of sea-surface currents using hf radar. Appl. Ocean Res. 158, 104542. http://dx.doi.org/10.1016/j.apor.2025.104542, https://www.sciencedirect.com/science/article/pii/S0141118725001300.

Cavaleri, L., Sclavo, M., Bertotti, L., 2010. Wave modelling in an ensemble prediction framework: Status and prospects. Ocean. Dyn. 60 (6), 1303–1310.

Delle Monache, L., Hacker, J.P., Zhou, Y., Deng, X., Stull, R.B., 2006. Probabilistic aspects of meteorological and ozone regional ensemble forecasts. J. Geophys. Res.: Atmospheres 111 (D24), http://dx.doi.org/10.1029/2005JD006917, arXiv:https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2005JD006917.

Epstein, E.S., 1969. Stochastic dynamic prediction. Tellus 21, 739–759.

Evensen, G., 2003. The ensemble Kalman filter: Theoretical formulation and practical implementation. Ocean. Dyn. 53 (4), 343–367.

Ferrari, F., Cassola, F., Tuju, P., Mazzino, A., 2021. RANS and LES face to face for forecasting extreme precipitation events in the liguria region (northwestern italy). Atmos. Res. 259, 105654.

Ferry, N., Parent, L., Garric, G., Bricaud, C., Testut, C.-E., 2010. Mercator global ocean reanalysis: Description and validation. Ocean. Sci. 6 (1), 151–166.

Global ocean physics analysis and forecast, E.U., Copernicus Marine Service Information (Cmems). Marine Data Store, http://dx.doi.org/10.48670/moi-00016.

Gneiting, T., 2014. Calibration of medium-range weather forecasts. In: European Centre for Medium-Range Weather Forecasts (ECMWF). Tech. rep., https://www.ecmwf.int/sites/default/files/elibrary/2014/9607-calibration-medium-range-weather-forecasts.pdf.

Gneiting, T., Raftery, A.E., 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. Mon. Weather Rev. 133 (5), 1098–1118.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. J. Amer. Statist. Assoc. 102 (477), 359–378. http://dx.doi.org/10.1198/016214506000001437.

Hamill, T.M., 2001. Interpretation of rank histograms for verifying ensemble forecasts. Mon. Weather Rev. 129 (3), 550–560.

Hamill, T.M., Whitaker, J.S., 2006. Probabilistic quantitative precipitation forecasts based on reforecast analogues: Theory and application. Mon. Weather Rev. 134, 3209–3229.

Hauswirth, S.M., Bierkens, M.F.P., Beijk, V., Wanders, N., 2023. The suitability of a seasonal ensemble hybrid framework including data-driven approaches for hydrological forecasting. Hydrol. Earth Syst. Sci. 27, 501. http://dx.doi.org/10.5194/hess-27-501-2023.

Houtekamer, P.L., Zhang, F., 2016. Review of the ensemble Kalman filter for atmospheric data assimilation. Mon. Weather Rev. 144 (12), 4489–4532.

Ingleby, B., 2015. Global assimilation of air temperature, humidity, wind and pressure from surface stations. Q. J. R. Meteorol. Soc. 141 (687), 504–517.

Lagomarsino-Oneto, D., Meanti, G., Pagliana, N., Verri, A., Mazzino, A., Rosasco, L., Seminara, A., 2023. Physics informed machine learning for wind speed prediction. Energy 268, 126628.

Lee, D.-J., Cho, S.-G., Yoo, C., Im, J., Yoon, D., Cha, D.-H., 2022. A novel ensemble learning for post-processing of NWP model's next-day maximum air temperature forecast in summer using deep learning and statistical approaches. Weather. Clim. Extrem. 37, 100410. http://dx.doi.org/10.1016/j.wace.2022.100410.

Lee Rodgers, J., Nicewander, W.A., 1988. Thirteen ways to look at the correlation coefficient. THe AMerican STatistician 42 (1), 59–66.

Leith, C.E., 1974. Theoretical skill of Monte-Carlo forecasts. Mon. Weather Rev. 102, 409–418.

Leutbecher, M., Palmer, T.N., forecasting, Ensemble., 2008. J. Comput. Phys. 227 (7), 3515–3539.

Mogensen, K., Bell, M., 2012. Ocean modelling and the seamless approach to weather and climate prediction. Q. J. R. Meteorol. Soc. 138 (665), 675–686. http://dx.doi.org/10.1002/qj.1971.

Palmer, T.N., 2019. Stochastic weather and climate models. Nat. Rev. Phys. 1 (6), 463–471. http://dx.doi.org/10.1038/s42254-019-0062-2.

Palmer, T.N., Č. Branković, Richardson, D., 2000. A probability and decision-model analysis of PROVOST seasonal multi-model ensemble integrations. Q. J. R. Meteorol. Soc. 126 (567), 2013–2033.

Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian model averaging to calibrate forecast ensembles. Mon. Weather Rev. 133 (5), 1155–1174.

Rasp, S., Lerch, S., 2018. Neural networks for postprocessing ensemble weather forecasts. Mon. Weather Rev. 146 (11), 3885–3900.

Samal, N., Ashwin, R., Singhal, A., Jha, S.K., Robertson, D.E., 2023. Using a bayesian joint probability approach to improve the skill of medium-range forecasts of the indian summer monsoon rainfall. J. Hydrol.: Reg. Stud. 45, 101284.

Scher, S., Messori, G., 2019. Weather and climate forecasting with neural networks: Using convolutional networks to predict a single time step. J. Adv. Model. Earth Syst. 11 (1), 1846–1853.

Silva Filho, T., Song, H., Perello-Nieto, M., Santos-Rodriguez, R., Kull, M., Flach, P., 2023. Classifier calibration: a survey on how to assess and improve predicted class probabilities. Mach. Learn. 1–50.

Simmons, J.A., Splinter, K.D., 2022. A multi-model ensemble approach to coastal storm erosion prediction. Environ. Model. Softw. 150, 105356. http://dx.doi.org/10.1016/j.envsoft.2022.105356, https://www.sciencedirect.com/science/article/pii/S1364815222000627.

Skamarock, W.C., Klemp, J.B., Dudhia, J., Gill, D.O., Barker, D.M., Duda, M.G., Huang, X.-Y., Wang, W., Powers, J.G., et al., 2008. A description of the advanced research WRF version 3. NCAR Tech. Note 475 (125), 10–5065.

Stammer, D., Balmaseda, M.A., Heimbach, P., et al., 2016. The ocean reanalyses intercomparison project (ORA-IP). J. Oper. Ocean. 9 (S2), S1–S46.

Stouffer, R.J., Gregory, J.M., Hall, A., Andrews, T., Palmer, M.D., Sedláček, J., Hu, X., Armour, K.C., Taylor, M.A., Ghan, S.J., 2017. CMIP6: The next generation of climate model intercomparison. Geophys. Res. Lett. 44 (12), 6667–6676. http://dx.doi.org/10.1002/2017GL074377.

Thoppil, P.G., Frolov, S., Rowley, C.D., et al., 2021. Ensemble forecasting greatly expands the prediction horizon for ocean mesoscale variability. Commun. Earth Environ. 2 (89), http://dx.doi.org/10.1038/s43247-021-00151-5.

Tu, T., Wang, J., Wang, C., Liang, Z., Duan, K., 2024. Reconstructing long-term natural flows by ensemble machine learning. Environ. Model. Softw. 177, 106069. http://dx.doi.org/10.1016/j.envsoft.2024.106069, https://www.sciencedirect.com/science/article/pii/S1364815224001300.

Tuju, P., Ferrari, F., Casciaro, G., Mazzino, A., 2022. The added value of high-resolution downscaling of the ECMWF-EPS for extreme precipitation forecasting. Atmos. Res. 280, 106458. http://dx.doi.org/10.1016/j.atmosres.2022.106458, https://www.sciencedirect.com/science/article/pii/S0169809522004446.

Whitaker, J.S., Hamill, T.M., 2002. Ensemble data assimilation without perturbed observations. Mon. Weather Rev. 130 (7), 1913–1924.

Wilks, D.S., 2011. Statistical methods in the atmospheric sciences, 100, Academic Press.

Zhang, Z., Hong, W.-C., 2019. Electric load forecasting by complete ensemble empirical mode decomposition adaptive noise and support vector regression with quantum-based dragonfly algorithm. Nonlinear Dynam. 98 (2), 1107–1136.

Zhao, T., Wang, Q., Bennett, J.C., Robertson, D.E., Shao, Q., Zhao, J., 2015. Quantifying predictive uncertainty of streamflow forecasts based on a bayesian joint probability model. J. Hydrol. 528, 329–340.