Distance Learning for Analog Methods

Paul Platzer, Arthur Avenas, Bertrand Chapron, Lucas Drumetz, Alexis Mouche, Pierre Tandeo, Léo Vinour

Distance Learning for Analog Methods

Paul Platzer[a b], Arthur Avenas[a b c], Bertrand Chapron[a b], Lucas Drumetz[b d], Alexis Mouche[a], Pierre Tandeo[b d e], Léo Vinour[a f]

[a] *Laboratoire d'Océanographie Physique et Spatiale, Univ. Brest/Ifremer/CNRS/IRD, F-29280 Plouzané, France*

[b] *Odyssey, Inria/IMT/CNRS, F-29280 Plouzané, France*

[c] *European Space Agency, Esrin, Italy*

[d] *IMT Atlantique, Lab-STICC, UMR CNRS 6285, 29238, Brest, France*

[e] *RIKEN Cluster for Pioneering Research, Kobe, 650-0047, Japan*

[f] *France Énergies Marines, Plouzané, France*

Corresponding author: Paul Platzer, paul.platzer@ifremer.fr

ABSTRACT: Analogs are similar states of a system, occurring at remote times within independent numerical simulations or previous observations. This concept has been developed in atmospheric sciences, and was further used in atmospheric and ocean sciences for forecasting, downscaling, upscaling, extreme event attribution, and many other applications. The distance used to find and rate analogs is a key feature of analog methods. Most studies are based on the Euclidean distance or other pre-defined metrics. In this investigation, we leverage distance learning algorithms originally designed for classification and regression and adapt them for statistical forecasting objectives, using in particular the continuous-ranked probability score as a loss function. Our algorithm allows to jointly optimize three key hyperparameters of analog methods: the feature space, the distance, and the number of analogs used. In particular, this algorithm allows to reduce the feature space dimension while keeping analog ensemble performances as high as possible, a key requirement for small and medium-sized datasets. We test our algorithm on an idealized chaotic system and on a small-size tropical cyclone dataset from meteorological agencies. Our experiments suggest that the optimal distance strongly depends on the forecast horizon and the number of available data, and that our algorithm allows for reasonable performances of analog ensemble methods even for small-size datasets. Our approach is not limited to forecasting and can assist the search for optimal hyperparameters of any analog method, enhancing exploration possibilities and improving overall performances.

SIGNIFICANCE STATEMENT: "History repeats itself." Today's weather is likely to be reminiscent of past, already observed weather. This is the notion of "analog" weather. Analogs were introduced to study the atmosphere, but were also used recently to study the ocean. They are used in many applications, including forecasts, or to estimate whether extreme events are influenced by climate change or not. To decide whether two distant-time images of the atmosphere (or ocean) are "analogs" of each other, one must define a similarity criterion, the "distance". The definition of the distance depends on the chosen application. This research aims at providing an algorithm to systematically tune the distance used in the definition of analogs, depending on the application.

1. Introduction

Analog methods rely on the search for neighbours of any given query in a database, also called "catalog" or "library", such as a reanalysis (*e.g.* ERA5, Hersbach et al. 2020) or an ensemble simulation (*e.g.* CMIP5, Taylor et al. 2012). They have been used in a wide range of applications in atmosphere and ocean sciences, including downscaling (Zorita and Von Storch 1999), upscaling (Yiou et al. 2014), ensemble forecasts (Delle Monache et al. 2013; Yiou 2014), tropical cyclone forecasting (Neumann and Hope 1972), extreme events detection and attribution (Jézéquel et al. 2018), importance sampling (Yiou and Jézéquel 2020), data assimilation (Tandeo et al. 2015; Lguensat et al. 2017), and interpolation (Zhen et al. 2020).

In atmospheric science, the idea of looking in observational archives to find similar states dates back to the concept of "points of symmetry" used in weather forecasting at least since Weickmann (1924) and then by Krick (1942) and Elliott (1943). The term "analogs" can be attributed to the study of Lorenz (1969) on atmospheric predictability. This term refers to methods and concepts developed in atmospheric science and now also used in ocean science (*e.g.* Le Bras et al. 2024), while other terms such as "neighbours" are more general and may refer to other concepts from dynamical systems (Lucarini et al. 2016) or machine learning (Peterson 2009). In this paper, "analog methods" explicitly refers to the use of neighbours-based methods in the specific context of atmospheric and ocean science.

The popularity of analog methods is probably due to their simplicity of interpretation and implementation. One simply needs to define a feature space, a distance (and kernel if weighting is applied), and to choose a number of analogs to be used. Then, one can perform any statistical task (*e.g.* averaging, linear regression, ensemble forecast) on the analog ensemble. These 3 key hyperparameters of analog methods (feature space, distance, and number of analogs used) are often directly chosen by the user depending on the application. Therefore, experience with using analog methods is required to make suitable hyperparameters choices, which prohibits an even wider use of analog methods in the atmospheric and ocean science community. Moreover, the sensitivity of analog methods to such choices is usually little or not explored unless it is the sole purpose of the study.

Empirical rules for the choice of these hyperparameters are as follows. First, the choice of features is based on the user's understanding of the physical system and on the objective task. Knowledge of dimensionality issues for analog methods in the case of limited-size datasets (Van den Dool 1994; Nicolis 1998; Platzer et al. 2021a) encourages one to reduce the number of features used to define analogs. Empirical orthogonal functions (EOFs, see *e.g.* Lorenz 1956) are commonly chosen as a dimension reduction technique to be applied before using analog methods (*e.g.* Benestad 2010). However EOFs were not developed for the purpose of analog methods, and are therefore likely to be sub-optimal.

Second, once features have been defined, one has to choose a distance between many available options (see, *e.g.*, the ones explored by Toth 1991; Matulla et al. 2008). Note that even after choosing a distance family (*e.g.* Euclidian l_2 vs. Manhattan l_1), an infinite number of variations of distance definitions are possible within each family, for instance by giving different weights to each feature, or by performing any one-to-one transformation of the feature vectors before computing the distance. This allows for a lot of explorations in the definition of the distance, so much that one usually cannot afford to perform by hand. As a consequence, the Euclidean distance in its simplest form is usually chosen for simplicity of implementation. Note also that the choice of distance only matters in the case of finite-sized datasets, as all distance functions[1] are equivalent in finite dimension (see, *e.g.*, appendix A in Platzer et al. 2021a). Therefore, the choice of distance is expected to be more important when using datasets of small or medium size.

[1] A distance function "dist$(\cdot,\cdot)$" should be positive ($\text{dist}[x,x'] > 0$), definite ($\text{dist}[x,x'] = 0$ iff. $x = x'$), symmetric ($\text{dist}[x,x'] = \text{dist}[x',x]$), and satisfy the triangle inequality ($\text{dist}[x,x''] \leq \text{dist}[x,x'] + \text{dist}[x',x'']$).

Third and finally, a near-optimal value of number of analogs is generally searched by trial-and-error, typically between $O(10) - O(200)$. If one requires a probabilistic estimation with an ensemble of analogs, one will need a large number of analogs, while for a deterministic estimation one analog can be enough in theory. However, even for deterministic estimates, the bias-variance trade-off rule (Hastie 2009) encourages one to use several analogs. A rule-of-thumb can be used by setting an upper-bound for the typical analog-to-target distance. Indeed, Platzer et al. (2021a) showed that the distance r_k to the k-th analog (ranked by growing distance) scales as $r_k \sim (k/N)^{1/d}$ where N is the number of independent samples in the catalog, and d is the local dimension, the latter being upper-bounded by p, the number of features in the distance definition. Setting the condition $r_k < 0.2$ (the value 0.2 is arbitrary), we have a sufficient rule for k which is $k < N \times 0.2^p$. Using such a rule for the maximum number of analogs to be used allows to have analog-to-target distances that do not exceed roughly 20% of the typical distance between two points chosen randomly in the dataset. However, this is only an approximate statistical upper-bound, and finding the optimal number of analogs for a given objective task usually requires a lot of testing efforts.

A typical meteorological system that was extensively studied with analog methods is the tropical cyclone (TC). In particular, analog methods have been used for the forecast of tropical cyclone tracks (*e.g.* Fraedrich et al. 2003; Langmack et al. 2012; Bonnardot et al. 2019; Bessafi et al. 2002) and intensities (*e.g.* Fetanat et al. 2013; Elsberry and Tsai 2014; Tsai and Elsberry 2014; Chen et al. 2016; Alessandrini et al. 2018; Tsai and Elsberry 2019; Bonnardot et al. 2019; Lewis et al. 2021). The tropical cyclone dynamical parameters are compiled in global databases called *best-tracks* (IBTrACS, Knapp et al. 2010). However, because of their underlying extreme ocean and atmospheric states, the crucial physical parameters of these phenomenon are still today largely undersampled in space and time. Recent studies also suggested that high-resolution observational data, *e.g* from satellite synthetic aperture radar, is required when studying this dynamical system (Avenas et al. 2023, 2024b). Thanks to new satellite missions and new observation strategies along with increased performances of the sensors, both the spatio-temporal resolution (Jackson et al. 2021) and the quality (Combot et al. 2020) of these data to capture the storm wind structure are greatly increasing during these years, soon providing a dataset of a reasonable size, although still limited, to be combined with statistical approaches such as analogs. However, such methods

will face the issues of small-size dataset and dimensionality mentioned earlier, and will therefore require a fine-tuning of the number of features and of the distance used.

Several authors have developed methodologies for the optimization of analog methods' hyperparameters. The kernels of Zhao and Giannakis (2016) are adapted to the local dynamics of the system under study and therefore provide optimal forecasts in the limit of large catalog size, as later demonstrated by Alexander and Giannakis (2020) using reproducing kernel Hilbert space theory. However, these methods do not tackle explicitly dimension reduction and small-size dataset issues. Also, they are targeted at forecasting of dynamical systems, while many other tasks can be performed with analog ensemble, and also require the optimisation of hyperparameters. McDermott and Wikle (2016) have given a Bayesian formulation of analog forecasting, allowing for optimisation of parameters through log-likelihood minimization. This approach yields a precise probabilistic formulation, and is flexible. Using a similar approach, Horton et al. (2017) perform optimization of analog methods through genetic algorithms, which allows to find global optimum and search through a wide range and number of hyperparameters. However, these two last methods require a thorough definition of hyperparameters to be optimised, while a more unifying framework would allow for a simpler representation and algorithmic implementation of the optimization of analog methods' hyperparameters.

Some authors have especially focused on the sensitivity of analog methods to the choice of distance. Toth (1991) compared nine different distances for the purpose of forecasting 700mb geopotential fields, and found some superiority for the root-mean-square difference between gradients of geopotential. Matulla et al. (2008) did a similar exercise for the purpose of precipitation downscaling, using five different distances and time-delayed embeddings in the space of EOFs. Authors jointly studied the effect of truncation in EOF space, the length of time-embedding, and the choice of distance. These studies allow for a detailed analysis of the performance of each distance choice for a specified task. However, as pointed by the authors of these two studies, they cannot generalize to other tasks than the one studied (forecasting for the first study and downscaling for the second). Moreover, one can generally not afford to conduct such tedious studies when designing a given analog method.

Other authors have introduced advanced methodologies for optimizing distances used in analog methods although this was not the direct topic of the study. This is the case of the tropical-cyclone

intensity analog forecasting scheme of Alessandrini et al. (2018). To find the best variables in more than 60 possible choices, the authors first searched for the best variable used alone, and then added new variables one-by-one, performing grid-search to find the best added variable and to define the respective weights of each new variable in the definition of the distance. Although it may prove efficient, the final solution given by such an algorithm is likely to be sub-optimal. Indeed, the fact that one predictor variable performs best alone does not guarantee that it should be retained when using several variables altogether. Furthermore, grid-search is a computationally intensive method to search for an optimum.

Fraedrich and Rückert (1998) optimized the distance used in analog methods by acting iteratively on weights given to coordinates used in Euclidean distance. However, by defining their own optimization step-rule, which was not studied elsewhere in the literature, the authors deprive themselves from the wealth of knowledge available for other well-known methods such as gradient descent. Indeed, the latter has well-established convergence properties, efficient algorithms (*e.g.* Kingma and Ba 2014), and has been used extensively in other fields of research. Furthermore, the algorithm of Fraedrich and Rückert (1998) is limited to optimizing weights, and a natural generalization would be the optimization of any linear transformation of the feature vector. Finally, the algorithm of Fraedrich and Rückert (1998) does not allow to perform dimension reduction, which is especially important when using analog methods with limited-size datasets. On the contrary, gradient-descent algorithms include regularization techniques which can help perform dimension reduction through feature selection.

In this paper, we focus on optimizing the distance used in analog methods. In particular, we leverage advances made in the field of machine learning through what is called "distance learning" or "metric learning" (Bellet et al. 2022). To our knowledge, distance learning algorithms have not been used before in analog methods for atmospheric and ocean sciences. We modified the "Metric Learning for Kernel Regression" algorithm (Weinberger and Tesauro 2007) by fixing the number of neighbors (here called "analogs") and adding a regularization term, just as Yang et al. (2012) modified the "Neighborhood Component Analysis" algorithm (Goldberger et al. 2004). Also, we adapted these algorithms to allow for probabilistic estimates such as ensembles, widely used in the context of atmospheric and ocean sciences. To our knowledge, this is the first time that a metric learning algorithm is adapted to probabilistic forecasts. For this purpose, we replace

the mean squared error in the loss function by the continuous ranked probability score (CRPS, see Hersbach 2000), a score which is well suited for ensemble-based predictions such as the ones provided by analog methods. Most metric learning algorithms, including the one presented here, aim at optimizing a Mahalanobis-type distance, which can be understood as a Euclidean distance computed after normalization with a positive semi-definite matrix, or equivalently after applying a linear transformation to the data. Note that what is called here "Mahalanobis-type distance" is more general than the "Mahalanobis distance" *per se* in which the data is renormalized by its covariance matrix. Here, the normalization matrix (or, equivalently, the linear transformation) is not set in advance but must rather be optimized.

Note that optimizing the distance *includes* optimizing the feature space: our algorithm allows to find optimal linear combination of any set of features, imposing sparsity if needed to perform dimension reduction, possibly removing irrelevant features. Furthermore, by adjusting the scale of the distance, we are able to adjust the bandwidth of analog methods, and therefore the number of analogs used. By adapting a simple, well-known algorithm from distance learning, we are therefore able to jointly optimize three key hyperparameters of analog methods: the feature space, the distance, and the number of analogs used. By basing our algorithm on one simple, well-known gradient-descent rule, we allow it to be applied to any existing analog methodology, in a unifying framework that simplifies the search for hyperparameters of analog methods and enables better overall performances.

The three-variable dynamical system of Lorenz (1963) is widely used to mimic atmospheric and oceanic systems because it is chaotic. It was developed as a simplified, low-order model for atmospheric convection, by focusing on first-order large-scale Fourier modes. We will use it as a well-known toy model to test the effect of our distance optimization algorithm on analog forecasting, allowing to explore in particular the effect of forecast horizon and catalog size, as well as comparing the results of our lagorithm when using two different loss functions: the mean-squared error or the continuous-ranked probability score (CRPS).

We also build a sub-dataset based on IBTrACS tropical cyclone data, used here to show that the proposed algorithm allows to select relevant variables for intensity forecasting. First, this should demonstrate that our algorithm can be used on small-size datasets, which is an interesting properties for tropical cyclone (TC) studies. Second, this should show that our algorithm allows

to perform dimension reduction by removing variables, which is a very important requirement for analog methods, as they are sensitive to the curse of dimensionality.

This paper is organized as follows. Basics of analog methods are recalled in section 2, before introducing the proposed algorithm and associated gradients. Experiments on the Lorenz system and on IBTrACS tropical cyclone data are reported in section 3 and 4. Conclusion and perspectives are drawn in section 5.

2. Algorithm

a. Analog methods

We assume that analogs of a query (or target) $\mathbf{x}$ are sought for in a catalog $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ of N independent d-dimensional vectors. The vectors $\mathbf{x}_j$ can be, for instance, the values of any variable (temperature, geopotential height) on a lon-lat grid, at a given time. The idea of analog methods is to search for analogues of a "common" situation $\mathbf{x}$, in order to predict the associated output variable $\mathbf{y}$, such as the precipitation at a given station.

In general, analogs are given weights which are increasing functions of the similarity between $\mathbf{x}$ and $\mathbf{x}_j$. Therefore, a statistical estimate of $\mathbf{y}$ is given by the weighted ensemble of values $\{\mathbf{y}_j,\, j \in I(\mathbf{x})\}$ associated with the analogs. From there on, one can use either pure empirical ensemble prediction (Yiou 2014), or sample average and standard deviation assuming a Gaussian probability distribution for $\mathbf{y}$, or weighted linear regression computed on the analog sample $\{(\mathbf{x}_j, \mathbf{y}_j),\, j \in I(\mathbf{x})\}$ (Platzer et al. 2021b).

b. Gradient of MSE of analog ensemble average

Our algorithm finds an optimal linear transformation of explanatory variables $\mathbf{x}$ before searching for analogs. The advantage of this algorithm is the simple computation of the gradient of the loss function. This is made possible by the use of Gaussian weights and a Mahalanobis-like distance. In our case, we will assume that the weights are of the following form:

$$p(\mathbf{x}_j|\mathbf{x}) \propto \exp\left(-\|\mathbf{A}(\mathbf{x}-\mathbf{x}_j)\|^2\right) \tag{1}$$

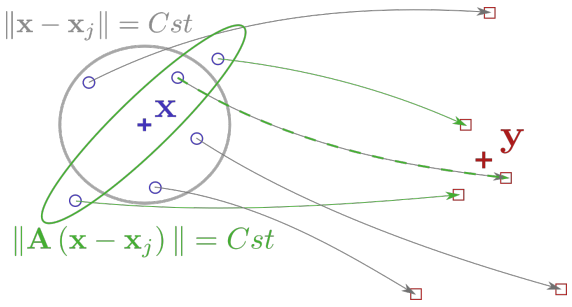

FIG. 1. Illustration of the process of making a linear transformation before applying analog ensemble methods. On the left are shown the input variables (also called features), with target value $\mathbf{x}$ (blue cross) surrounded by potential analogs $\mathbf{x}_j$ (blue circles). On the right are shown the output variables, with true output $\mathbf{y}$ (red cross) and outputs $\mathbf{y}_j$ (red squares) associated with analogs $\mathbf{x}_j$. The selection of analogs is performed in input space (or feature space), where the grey circle corresponds to a level-curve of constant Euclidean distance to the target value $\mathbf{x}$, and the green circle corresponds to constant distance to $\mathbf{x}$ after applying a linear transformation $\mathbf{A}$.

where $\mathbf{A}$ is a $p \times d$-matrix with $p \leq d$, and $||\cdot||$ is the Euclidean norm. After normalization over the set of analogs $I(\mathbf{x})$, we find that each analog is given the empirical probability:

$$p(\mathbf{x}_j|\mathbf{x}) = \frac{\exp\left(-\|\mathbf{A}(\mathbf{x}-\mathbf{x}_j)\|^2\right)}{\sum_{k\in I(\mathbf{x})}\exp\left(-\|\mathbf{A}(\mathbf{x}-\mathbf{x}_k)\|^2\right)}\,. \tag{2}$$

The procedure of applying a linear transformation before selecting and weighing analogs is illustrated in Fig. 1. In this example, applying the matrix $\mathbf{A}$ allows to select analogs $\mathbf{x}_j$ with associated outputs $\mathbf{y}_j$ that are closer to the true output $\mathbf{y}$, compared to the result of using the original distances (with $\mathbf{A}$ replaced by the identity matrix).

The objective is now to find the optimal $p \times d$-matrix $\mathbf{A}$, which is a linear transformation of vectors $\mathbf{x}$ to a space in which the analogs are sought for using the simple Euclidean distance. This is equivalent to finding the optimal Mahalanobis-type distance $\mathrm{dist}(\mathbf{x}, \mathbf{x}_j) = (\mathbf{x}-\mathbf{x}_j)^T\mathbf{A}^T\mathbf{A}(\mathbf{x}-\mathbf{x}_j)$ (where T-subscript stands for "transpose") with positive symmetric matrix $\mathbf{A}^T\mathbf{A}$.

It should be stressed again here that our use of the expression "Mahalanobis-type distance" differs from another common use, where the symmetric, positive, semi-definite matrix $\mathbf{A}^T\mathbf{A}$ is replaced by the covariance of the data (McLachlan 1999), which would here be the catalog. Instead, we optimize the matrix $\mathbf{A}$, and therefore we say that we optimize the distance within the family of Mahalanobis-type distances. Note that some distance learning algorithms choose to directly

estimate the symmetric, positive, semi-definite matrix, which can sometimes allow to define convex optimization algorithms (Globerson and Roweis 2005). However, optimizing $\mathbf{A}$ has the advantage to ease the interpretation, in particular in the context of dimension reduction.

To keep the analogy with the algorithm of Goldberger et al. (2004), we optimize the mean-square error (MSE) of the analog prediction from the catalog, using a leave-one-out procedure. We compute the square error of analog average prediction of $\mathbf{y}_i$ from the truncated catalog $\{\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_N\}$. For simplicity, we note $I(i) = I(\mathbf{x}_i)$ the indices of analogs of $\mathbf{x}_i$ in the truncated catalog, and $p(j|i) = p(\mathbf{x}_j|\mathbf{x}_i)$ is the probability given to value $\mathbf{y}_j$ associated with analog $\mathbf{x}_j$ of target $\mathbf{x}_i$ using the truncated catalog. Note that $p(j|i) \neq p(i|j)$ is not symmetric, because of the normalization factor in Eq. (2). The MSE can then be written:

$$\mathrm{MSE}(\mathbf{A}) = \frac{1}{N} \sum_{i=1}^{N} (\hat{\mathbf{y}}_i - \mathbf{y}_i)^T (\hat{\mathbf{y}}_i - \mathbf{y}_i) \,, \tag{3}$$

where we use the notation $\hat{\mathbf{y}}_i = \sum_{j \in I(i)} p(j|i) \mathbf{y}_j$ for the average analog prediction of $\mathbf{y}_i$. This MSE has the following gradient:

$$\frac{\partial \mathrm{MSE}}{\partial \mathbf{A}} = \frac{2}{N} \sum_{i=1}^{N} (\hat{\mathbf{y}}_i - \mathbf{y}_i)^T \sum_{k \in I(i)} \frac{\partial p(k|i)}{\partial \mathbf{A}} \mathbf{y}_k \,. \tag{4}$$

Note that what we call "gradient" for simplicity is actually a "sub-gradient", because we neglect the discontinuities of MSE($\mathbf{A}$) due to changes in the the lists of analog indices $I(i)$ with the change of $\mathbf{A}$. The computation of the sub-gradient however is simpler than the true gradient, and sufficient for convergence (Held et al. 1974). In the following we therefore only use the word "gradient".

Since we use exponential weights, the gradient can be computed easily. Indeed, the gradient of $p(k|i)$ with respect to matrix $\mathbf{A}$, which is a matrix of the same shape as $\mathbf{A}$, is given by:

$$\frac{\partial p(k|i)}{\partial \mathbf{A}} = \mathbf{A} p(k|i) \left(\sum_{l \in I(i)} p(l|i) \mathbf{x}_{il} \mathbf{x}_{il}{}^T - \mathbf{x}_{ik} \mathbf{x}_{ik}{}^T \right) , \tag{5}$$

where we use the notation $\mathbf{x}_{ik} = \mathbf{x}_k - \mathbf{x}_i$.

After simplifications, similar to the ones of Goldberger et al. (2004), the gradient of the MSE can be rewritten:

$$\frac{\partial \mathrm{MSE}}{\partial \mathbf{A}} = \mathbf{A}\frac{2}{N}\sum_{i=1}^{N}(\hat{\mathbf{y}}_i - \mathbf{y}_i)^T \sum_{j\in I(i)} p(j|i)(\hat{\mathbf{y}}_i - \mathbf{y}_j)\mathbf{x}_{ij}\mathbf{x}_{ij}{}^T \,. \tag{6}$$

where there are two independent products inside the sum: one *dot* product between vectors $(\hat{\mathbf{y}}_i - \mathbf{y}_i)$ and $(\hat{\mathbf{y}}_i - \mathbf{y}_j)$, and a *tensor* product between the vector $\mathbf{x}_{ij}$ and itself. Applying the matrix $\mathbf{A}$ of shape (p,d) to the tensor product of shape (d,d) gives a matrix of the same size as $\mathbf{A}$, which is the size of the gradient.

The advantage of expressing the gradient in this way is that many terms are already computed during the analog average prediction step. Therefore, to compute the gradient of the MSE on the catalog, one must simply apply matrix $\mathbf{A}$ to a product of partly pre-computed terms.

c. Gradient of CRPS of analog ensemble

In atmospheric and ocean science, estimation of uncertainty is a key feature of any estimation algorithm. Also, the strength of analog methods is the cheap generation of ensembles. Therefore, using the MSE to assess analog methods is too restrictive and we propose to use the continuous-ranked probability score (CRPS), widely used in atmospheric sciences (Hersbach 2000).

The CRPS of a one-dimensional probabilistic forecast with cumulative probability distribution F, compared to a true scalar outcome y, can be expressed as :

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} \left[F(y') - \mathbb{1}(y' > y)\right]^2 \mathrm{d}y' \,, \tag{7}$$

where $\mathbb{1}(y' > y)$ is the indicator function which equals zero when $y' <= y$ and equals one when $y' > y$. Another form of the CRPS which is more convenient to our purposes is the following :

$$\mathrm{CRPS}(F, y) = \mathbb{E}_F|Y - y| - \frac{1}{2}\mathbb{E}_F|Y - Y'| \,, \tag{8}$$

where $\mathbb{E}_F$ is the expectation and Y and Y' are random variables distributed according to F. The first term of this equation is the mean-absolute error of the forecast. The second-term is the negative half of the mean absolute difference between two variables distributed according to the forecast. When we have an ensemble forecast, this is the mean absolute difference between two forecast members. Using our notations, this allows to express easily the CRPS of the analog ensemble forecast of y_i from the values $\{y_j\}_{j\in I(i)}$ distributed according to $\{p(j|i)\}_{j\in I(i)}$.

$$\mathrm{CRPS}_i = \sum_{j \in I(i)} p(j|i) y_{ji} - \frac{1}{2} \sum_{j,k \in I(i)} p(j|i) p(k|i) y_{jk} \,, \tag{9}$$

where we use the short notation for difference of absolute values $y_{ji} := |y_j - y_i|$. This can be written more concisely as $\mathrm{CRPS}_i = \mathrm{MAE}_i - \frac{1}{2}\mathrm{MAD}_i$ with $\mathrm{MAE}_i := \sum_{j \in I(i)} p(j|i) y_{ji}$ and $\mathrm{MAD}_i := \sum_{j,k \in I(i)} p(j|i) p(k|i) y_{jk}$. Finally, we are interested in optimizing:

$$\overline{\mathrm{CRPS}} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{CRPS}_i \,. \tag{10}$$

Following similar steps as in the previous section, we find the following expression for the gradient :

$$\frac{\partial \overline{\mathrm{CRPS}}}{\partial \mathbf{A}} = \mathbf{A} \frac{1}{N} \sum_{i=1}^{N} \sum_{j \in I(i)} p(j|i) \left\{ \mathrm{MAE}_i - \mathrm{MAD}_i - \left(y_{ji} - \sum_{k \in I(i)} p(k|i) y_{kj} \right) \right\} \mathbf{x}_{ij} \mathbf{x}_{ij}^T \,. \tag{11}$$

This expression can be generalized to the case of vector-output $\mathbf{y}$ by performing a sum of CRPS over all coordinates.

If one is interested in computing the CRPS or of a multi-dimensional output, then one simple possibility is to use the average of CRPS over all coordinates. The gradient of the coordinate-averaged CRPS is then given by the above formula, replacing MAE_i, MAD_i and y_{ij} terms by their coordinate-average.

We provide modified formulas for the weighted-CRPS in appendix A, which could be used for applications in which specific values (such as extreme values) are of interest.

d. Algorithm

Goldberger et al. (2004) designed an algorithm which optimizes the matrix of a Mahalanobis-type distance for classification purposes in a smoothed version of a nearest neighbour algorithm. Yang et al. (2012) modified the algorithm of Goldberger et al. (2004), keeping only nearest neighbours for classification. Weinberger and Tesauro (2007) designed an algorithm similar to Goldberger et al. (2004), for optimizing a Mahalanobis-type distance for kernel regression. Here, we adapt the algorithm from Weinberger and Tesauro (2007) to the case of analog forecasting, by keeping only a finite number of nearest neighbours (analogs) in the computation, and possibly adding regularization terms. Therefore, our algorithm is a modified version of Weinberger and Tesauro

(2007) just as the algorithm of Yang et al. (2012) is a modified version of Goldberger et al. (2004). Furthermore, we adapt the algorithm to the case where the loss function is not a mean-square error, but a continuous ranked probability score, which is more suited to probabilistic ensemble-based estimators used in atmospheric and ocean sciences.

One may be interested in sparse representation of the matrix $\mathbf{A}$, for instance to perform feature selection as we demonstrate in section 4. To do so, we add the well-known l_1/l_2 regularization term to the loss function (Yin et al. 2014) :

$$\text{Loss}(\mathbf{A}) = \lambda \frac{\|\mathbf{A}\|_1}{\|\mathbf{A}\|_2} + \begin{cases} \text{MSE}(\mathbf{A}) \\ \text{CRPS}(\mathbf{A}) \end{cases} \tag{12}$$

with fixed parameter $\lambda > 0$, and $\|\mathbf{A}\|_1$ and $\|\mathbf{A}\|_2$ are the l_1 and l_2-norms of matrix $\mathbf{A}$, namely the sum of its coefficients' absolute values, and the square-root of the sum of its coefficients' squared values. This additional term has sub-gradient $\frac{\partial}{\partial \mathbf{A}} \frac{\|\mathbf{A}\|_1}{\|\mathbf{A}\|2} = \frac{\text{sign}(\mathbf{A})}{\|\mathbf{A}\|_2} - \frac{\|\mathbf{A}\|_1}{\|\mathbf{A}\|_2^3}\mathbf{A}$. This term has the advantage of being scale-invariant, which is a desired property of our algorithm because adjusting the scale of $\mathbf{A}$ allows to adjust the number of analogs used. If we had used a simple l_1 Lasso regularization (Tibshirani 1996) or l_2 Ridge regularization (Hoerl and Kennard 1970), then we would have biaised our algorithm towards low-scale $\mathbf{A}$, *i.e.* towards a high number of analogs used. However, this choice could be useful in applications where one wants to force the analog method to use a large number of analogs, and therefore reducing variance at the cost of raising bias.

Finally, the matrix $\mathbf{A}$ is updated at fixed learning rate $\alpha > 0$ with the gradient descent rule (Bottou 2012):

$$\mathbf{A}_{new} = \mathbf{A}_{old} - \alpha \frac{\partial \text{Loss}}{\partial \mathbf{A}} \tag{13}$$

Based on that updated matrix $\mathbf{A}_{new}$, one can compute the new probabilities $p(j|i)$ and associated MSE and gradient, and update again the matrix using Eq. (13).

The algorithm is summarized in Algorithm 1. In this algorithm, the search for k neighbors is denoted kNN. Particular cases of this algorithm include:

- *Weighting of coordinates*: by imposing $\mathbf{A}$ to be diagonal.
- *Dimensionality reduction*: by setting $\mathbf{A}$ to be of shape $p \times d$ with p small.

Note that when one imposes $\mathbf{A}$ to be diagonal, the tensor products $\mathbf{x}_{ij}{\mathbf{x}_{ij}}^T$ must be replaced by vectors of size d and whose coordinates are the square of the vector $\mathbf{x}_{ij}$ (this is the diagonal of the tensor $\mathbf{x}_{ij}{\mathbf{x}_{ij}}^T$). This has the advantage of begin much less computationally-intensive than $d \times d$ tensors.

Algorithm 1: Optimize Mahalanobis distance for Analog Prediction

Function `OptimizeMahalanobis(` $\mathbf{A}_0$, $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, $\{y_1, \dots, y_N\}$, α, $\lambda > 0$, n, k)

- $\mathbf{A} = \mathbf{A}_0$;
- **for** $t \in [1, n]$ **do**
  - Build tree for $\{\mathbf{A}\mathbf{x}_i\,,\, i \in [1, N]\}$
  - **for** $i \in [1, N]$ **do**
    - $I(i) := k\mathrm{NN}_j ||\mathbf{A}(\mathbf{x}_i - \mathbf{x}_j)||$;
    - $w_{ij} := \exp\left(-||\mathbf{A}(\mathbf{x}_i - \mathbf{x}_j)||^2\right)$;
    - $p(j|i) := \frac{w_{ij}}{\sum_{k \in I(i)} w_{ik}}$;
  - **end**
  - $\mathbf{G}$:= (6) or (11) or (A3)
  - $\mathbf{G} \leftarrow \mathbf{G} + \lambda \left(\frac{\mathrm{sign}(\mathbf{A})}{\|\mathbf{A}\|_2} - \frac{\|\mathbf{A}\|_1}{\|\mathbf{A}\|_2^3} \mathbf{A} \right)$
  - $\mathbf{A} \leftarrow \mathbf{A} - \alpha \mathbf{G}$;
  - $\mathrm{E}_t \leftarrow \begin{cases} \mathrm{MSE}(\mathbf{A}) \\ \mathrm{CRPS}(\mathbf{A}) \end{cases}$
- **end**
- **return** $\mathbf{A}$, $(\mathrm{E}_1, \dots, \mathrm{E}_n)$;

3. Lorenz System experiments

To analyze the behaviour of our algorithm, we first use the well-known chaotic, three-variable deterministic dynamical system of Lorenz (1963), in its usual setting, following the equations:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \sigma(y - x), \tag{14}$$

$$\frac{\mathrm{d}y}{\mathrm{d}t} = x(\rho - z) - y, \tag{15}$$

$$\frac{\mathrm{d}z}{\mathrm{d}t} = xy - \beta z, \tag{16}$$

with parameters $\sigma = 10$, $\beta = 8/3$ and $\rho = 28$. The equations are solved numerically using a fourth-order Runge-Kutta explicit scheme (Butcher 1996) with time-step 0.01 (non-dimensional time). This system of equation approximates the large-scale behaviour of atmospheric convection,

and bears properties of atmospheric and ocean circulation, in particular the sensitivity to initial condition.

Note that the classical notation (x, y, z) used here for the three coordinates of the system must not be confused with the notations $\mathbf{x}$ and $\mathbf{y}$ of the previous section, used to denote the predictor and predicted variables, respectively. In particular, this section will make the particular choice of $\mathbf{x} = [x(t), y(t), z(t)]$ (three coordinates as predictor) and either $\mathbf{y} = [x(t+h), y(t+h), z(t+h)]$ (forecast of all three variables at horizon h) or $\mathbf{y} = [z(t+h)]$ (forecast of the last variable at horizon h). For our numerical experiment, we choose to set $\mathbf{A}$ to be a 3x3 matrix, and therefore the transformed variables $\mathbf{Ax}$ are of the same number (3) as the original variables $\mathbf{x}$. Note that other choices could have been retained for the shape of $\mathbf{A}$, such as a 1x3 matrix to retain only one variable. We could also have included other input variables in the $\mathbf{x}$ vector by using delayed-coordinates (Sauer et al. 1991). The only constraint is that the number of columns in $\mathbf{A}$ equals the size of $\mathbf{x}$ (*i.e.* the number of input variables). We chose this simple 3x3 setting with the original coordinates of the Lorenz system as input variables for illustrative testing of our method.

In all the experiments involving the Lorenz system, we will use k =200 analogs for each forecast ensemble, that is $|I(\mathbf{x})|$= 200 for all $\mathbf{x}$ (using notations from section 2). The set of analogs will be defined by the 200 indices j of nearest neighbors of $\mathbf{x}$ according to the Euclidean distance $\|\mathbf{A}(\mathbf{x}_j - \mathbf{x})\|$. We will use no regularization term in the algorithm (*i.e.* $\lambda = 0$ in Eq. (12)), because it is unnecessary for this low-dimensional system.

The catalogs used in these experiments will be generated from long trajectories of the numerically-integrated Lorenz System. Elements of the catalog will comprise values of $\mathbf{x}_i = [x(t_i), y(t_i), z(t_i)]$ along with corresponding values of either $\mathbf{y}_i = [x(t_i + h), y(t_i + h), z(t_i + h)]$ or $\mathbf{y}_i = [z(t_i + h)]$ at horizon $h > 0$. The sequence of values $\{t_1 \ldots, t_N\}$ will be of the type $t_i = t_1 + (i-1)\Delta t$ where the time separating two elements, $\Delta t = 0.64$ (non-dimensional time), is chosen so that they can be considered independent. At least, this value of $\Delta t = 0.64$ is enough so that $x(t_{i+1})$ and $x(t_{i-1})$ are not part of the 200 analogs of $x(t_i)$ in our experiments.

Before applying our algorithm, we standardize each variable, that is we divide the values of x by $\left\langle (x - \langle x \rangle)^2 \right\rangle^{1/2}$, the values of y by $\left\langle (y - \langle y \rangle)^2 \right\rangle^{1/2}$, and the values of z by $\left\langle (z - \langle z \rangle)^2 \right\rangle^{1/2}$. Therefore the matrix $\mathbf{A}$ applies a transformation on standardized variables. For readability purposes, we will keep the notations x, y and z for the standardized variables.

We will first apply our algorithm to minimize the RMSE of the analog ensemble average prediction, and explore varying forecast horizons and forecasted variable, then we will investigate varying catalog sizes, and finally we will compare the results of minimizing the RMSE versus minimizing the CRPS of the analog forecast ensemble.

a. Varying forecast horizon and variable

In this subsection, we use a catalog of fixed size 10^5 generated from one long trajectory of the Lorenz System, and we vary the forecast horizon h while keeping constant the values $\{t_1, \ldots, t_N\}$ as defined earlier.

To begin with, we apply our algorithm to minimize the MSE of the analog forecast at horizon h =0.01 (non-dimensional time), which is the time-step at which we perform the Runge-Kutta integration of the Lorenz system's equations. Note that at this very small horizon, the ideal forecast is very close to a persistence forecast.

Our algorithm requires an initial value $\mathbf{A}_0$ for the matrix $\mathbf{A}$. Note that as we are solving a non-convex optimization problem, the choice of initial value influences the final result. We will start with a standard choice for $\mathbf{A}_0$ that we name "isotropic" ($\mathbf{A}_0 = \mathbf{A}_{iso}$), that is the identity matrix:

$$\mathbf{A}_{iso} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{17}$$

Therefore, our algorithm starts with a baseline of standardized variables. Then, to choose the learning rate α (Eq. 13), we first compute the initial mean-squared error MSE_0 of the analog ensemble average, using distances $\|\mathbf{A}_0(\mathbf{x}_i - \mathbf{x}_j)\|$ where $\mathbf{A}_0$ is set to $\mathbf{A}_{iso}$, and the leave-one-out methodology to compute the error, *i.e.* we make a forecast of $\mathbf{y}_i$ using all the catalog without the element $(\mathbf{x}_i, \mathbf{y}_i)$. Finally, we set the learning rate to be $\frac{60}{\mathrm{MSE}_0}$, and we run our algorithm through n =60 iterations, after which the algorithm is converged to a value that we note $\mathbf{A}_{con}(0.01)$ where "con" stands for "converged". Note that MSE_0 depends on whether we are learning a matrix for the forecast of the z-variable or of all three variables (x, y, z). Fig. 2(a,b,d,e) allows to compare the distances $\|\mathbf{A}_{iso}(\mathbf{x}_i - \mathbf{x}_j)\|$ (Fig. 2.a,d) to the distances obtained after 60 iterations of our algorithm for the forecast of the z-variable at horizon h =0.01 (Fig. 2.b,e). One can see that the optimized

distances are larger than the initial ones, this means that the probabilities given to the analogs $p(i|j)$ are sharper: the selection of analogs is narrower. Also, one can witness a change in the relative weights given to each coordinate: the ellipsis around the red dot has rotated, so that distances grow faster with coordinate z in Fig. 2(e) than in Fig. 2(d) compared to how they grow with coordinate x. This shows that the optimal distance is anisotropic.

Then, the algorithm is run again, but at horizon $h = 0.02$, and starting from the previously converged value $\mathbf{A}_0 = \mathbf{A}_{con}(0.01)$. Again, we compute the initial MSE at horizon h =0.02, and this time we set $\alpha = \frac{30}{\mathrm{MSE}_0}$ and run only n =20 iterations which is enough for the algorithm to converge. This allows the algorithm to move to the closest local minimum of MSE when changing slightly the forecast horizon. We then repeat this operation, raising the horizon by 0.01, running the optimization with n =20 iterations after updating the value of $\mathbf{A}_0$ and recomputing $\alpha = \frac{30}{\mathrm{MSE}_0}$ at each new horizon. Note that one succession of optimizations is run for the forecast of the variable z and the other succession of optimizations is run independently for the forecast of the three variables (x, y, z). The result of running this procedure until horizon $h = 0.32$ is shown in Fig. 2(c,f). Comparing with Fig. 2(a,b,d,e) again shows a change in the orientation of the ellipsis of constant distances, so that in particular the relative weights of each original coordinate in the modified distance have changed. Also, all distances are larger indicating an even more selective choice of analogs (*i.e.* even sharper distributions $p(i|j)$).

Fig. 3(a,b) shows the evolution of all coefficients of the optimized $\mathbf{A}_{conv}$ with forecast horizon, and depending on the choice of either $\mathbf{y}_i = [z(t_i + h)]$ or $\mathbf{y}_i = [x(t_i + h), y(t_i + h), z(t_i + h)]$. To aid the interpretation, the l_2-norms of $\mathbf{A}$'s rows are also shown (*i.e.* $\sqrt{A_{xj}^2 + A_{yj}^2 + A_{zj}^2}$ for $j \in \{x, y, z\}$), which allow to assess the relative importance of each original coordinate in the converged optimal distance. This shows in particular that for small forecast horizons ($h < 0.05$) the z-variable is the most important variable when one is concerned with a forecast of the z-variable itself, which was expected. Also, one can notice the growth of coefficients (and thus of the overall norm of $\mathbf{A}$) with horizon: this shows that a more selective choice of analogs must be done when concerned with a forecast at large horizon, which is also expected due to the chaotic nature of the Lorenz system. Note that the growth of the norm $\mathbf{A}$'s rows with horizon concerns the x and y coordinates. We attribute this observation to the fact that these two coordinates are indicative of which "wing" one is in (see Fig. 2): each "wing" corresponds to one of the two unstable fixed points around which

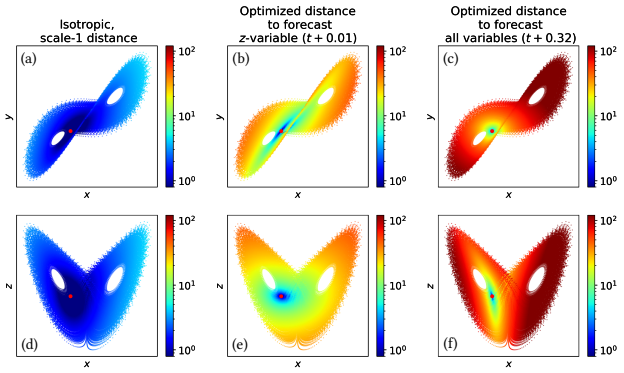

FIG. 2. Visualization of distances to a random point (red) across the Lorenz attractor, comparing (a,d) unoptimized, isotropic distance with (b,e) distances optimized to minimize the RMSE of the mean analog forecast of z-variable at short horizon and (c,f) mean analog forecast of all three state variables at large horizon. The chosen forecast horizon can be interpreted as $\sim$ 1 hour and $\sim$ 1 day in atmospheric time scale.

solutions of the Lorenz system orbit before an eventual transition to the other wing. Finally, one can see that even at large forecast horizon, the optimal distance found when trying to minimize the RMSE of the forecast of only one variable differs from the optimal distance that is found when trying to forecast all three variables of the Lorenz System. This indicates that there is no universal distance that could be used for any analog method. Fig. 3(e,f) shows that the gain in RMSE after optimization when compared to the RMSE found with $\mathbf{A}_{iso}$ is constant throughout the range of forecast horizons.

b. Varying catalog size

In this section, we will build catalogs of different sizes using the same long trajectory of $6{\times}10^6$ non-dimensional time steps. From this trajectory, we will extract catalogs of different sizes, and for each catalog size we extract 10 different catalogs to get variability in the results when fixing only the catalog size. To do so, we first take regular sampling times $\{t_1^*, \ldots, t_{10N}^*\}$ of the whole trajectory, separated by $t_{i+1}^* - t_i^* = \frac{t_{10N}^* - t_1^*}{10N\Delta t}$ where N is the desired catalog size, $\Delta t = 0.64$ (non-dimensional time), and t_1^* and t_{10N}^* are the first and last time of the whole trajectory. Then, we take a permutation Perm: $[1, 10N] \mapsto [1, 10N]$, and build 10 catalogs from indices $\{t_{\text{Perm}(1)}, \ldots, t_{\text{Perm}(N)}\}$, $\{t_{\text{Perm}(N+1)}, \ldots, t_{\text{Perm}(2N)}\}$, $\{t_{\text{Perm}(2N+1)}, \ldots, t_{\text{Perm}(3N)}\}$, up to $\{t_{\text{Perm}(9N+1)}, \ldots, t_{\text{Perm}(10N)}\}$. This procedure allows to generate 10 catalogs from a single long trajectory for each chosen catalog size

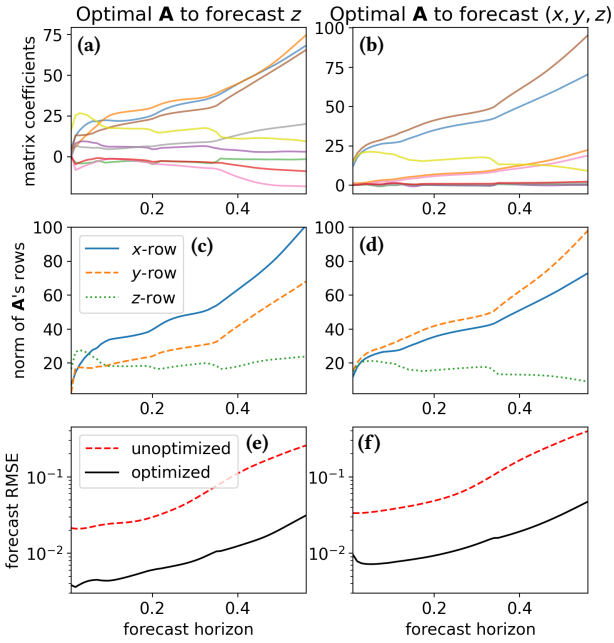

FIG. 3. (a,b) Coefficients of matrix **A** (see Fig. 4 for the color-to-coefficient correspondence) versus forecast horizon. (c,d) Norm of rows of transform matrix **A** versus forecast horizon. (e,f) Analog forecast RMSE versus forecast horizon. (a,c,e) Forecast of z-variable. (b,d,f) Forecast of whole state-space vector.

N. We use this procedure for 30 values of the catalog size N in the range $N \in [4\times10^3, 4\times10^5]$ with regular sampling in log-scale within this interval.

In this experiment, we run our algorithm for the objective of minimizing the RMSE of the forecast of all three variables at horizon $h = 0.32$ (non-dimensional time). The algorithm was run for each catalog constructed as explained above, with the linear transformation always initialized with the identity matrix as $\mathbf{A}_0 = \mathbf{A}_{iso}$. The algorithm is run for 60 iterations, with a learning rate set to $\frac{50}{\text{MSE}_0}$ (which depends on the catalog used) similarly to the previous experiment. The results of this experiment are shown in Fig. 4. The ratio between optimized RMSE (with the final value of **A**) and unoptimized RMSE (with $\mathbf{A}_{iso}$) is nearly constant throughout the tested values of catalog size. The values of coefficients of the final optimized matrix **A** show a very strong dependency with catalog size for small catalog sizes (below 10^4), also with greater variability of the coefficients for a fixed catalog size. For larger values of the catalog size, the dependency of coefficients with catalog size is much weaker, as well as the variability within a given value of catalog size. In particular, for catalog sizes below $\sim 10^4$, the diagonal coefficients acting on the the x (blue lines) and z (yellow

lines) variable are decreasing functions of catalog size, suggesting the need for a more rigorous selection of analogs when the catalog size is small. We interpret this result as the consequence of the fact that when the catalog size is small, the chance to find poor quality analogs in the 200 nearest neighbors is higher, and therefore analog selection must be more meticulous. Moreover, this behaviour is not observed for the diagonal coefficient acting on y, which suggests that when changing the catalog size, one must not only adapt the number of analogs used, but also the very definition of the distance, such as the importance given to specific variables (or features) in the definition of the distance.

A striking property of Fig. 4(b) is the change of slope of the almost convex black curve of optimized RMSE around catalog size $\sim 10^4$. In particular, for catalog sizes above 10^4, the black curve has a flatter slope than the red curve. This convexity of the black curve indicates that our optimization allows to approach faster the properties of an analog ensemble with large catalog size. In other words, the necessity to have a large catalog is less critical when optimizing the properties of analog methods. Our algorithm thus allows to (moderately) compensate for small-sized datasets.

Note that the behaviour observed for low values of the catalog size shows signs of the beginning of overfitting: coefficients are becoming large and highly variable. However, to check whether we are truly witnessing overfitting, we have computed the RMSE of the analog ensemble forecasts using the same catalogs as in the previous experiment, the same coefficients as fitted on these catalogs using our algorithm, but making forecasts on an independently generated test set of 10^4 elements from a long trajectory, with elements separated by 0.64 non-dimensional time intervals. This independent test gives indistinguishingly (not shown) the same values as the RMSE shown in Fig. 4(b) and computed on the catalog (*i.e.* on the training set with leave-one-out methodology). This shows that the high values of the coefficients (1,1) and (3,3) of the fitted matrix $\mathbf{A}$ really help to improve the performances of analog forecasting for small catalog size.

c. Minimizing CRPS vs. minimizing RMSE

Finally, we investigate the difference in the results of our algorithm when used with the same catalog, the same forecast objective (forecast of the z variable at horizon $h = 0.04$), but with different loss function, either the RMSE or the CRPS.

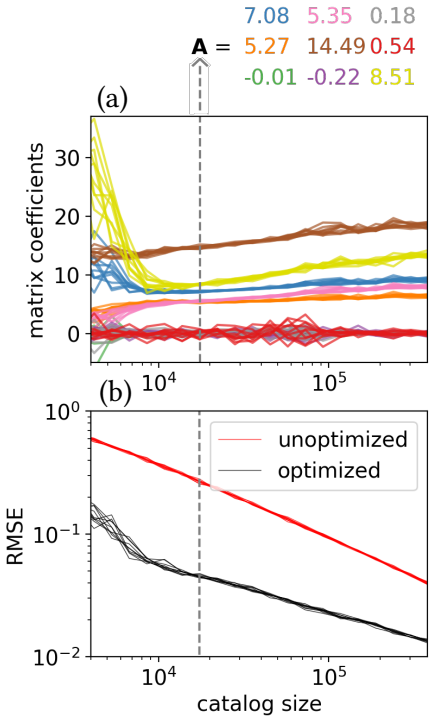

FIG. 4. Matrix coefficients (a) of optimized linear transformation **A** for average analog forecast of the Lorenz system at horizon 0.32 (1 day), as a function of catalog size. RMSE is also shown (b) for both optimized (black) and unoptimized (red) analog forecast. An example of optimized matrix **A** is shown on top for catalog size 10 826. For each catalog size, 10 optimizations are run for 10 independent catalogs to account for variability in the optimization process. Optimization is run at constant learning rate and fixed number of iterations, initialized with the identity matrix. 200 analogs are retained for forecast.

We use the same catalog as in section **3.**a), and we take the result of the experiment described in the same section to obtain the matrix **A** optimized for the RMSE of the average analog forecast of the z variable at horizon h =0.04. To compare with the results for the CRPS, we use a similar procedure to find the optimized matrix **A**, starting by running our algorithm to minimize the CRPS of the forecast of the z-variable at horizon 0.01, with $\mathbf{A}_0 = \mathbf{A}_{iso}$, 200 iterations, and a learning rate of $\frac{10^3}{\mathrm{CRPS}_0}$, where CRPS_0 is defined analogously to MSE_0 using the initial value of $\mathbf{A}_0$. Then, we optimize the CRPS for the forecast of the z-variable ot horizon $h = 0.02$ initializing the matrix with the previous optimized value $\mathbf{A}_0 = \mathbf{A}_{con}(0.01)$ is was done for the RMSE. The algorithm is run for 100 iterations at rate $\frac{1.5\times10^3}{\mathrm{CRPS}_0}$. The operation is repeated to go from horizon 0.02 to 0.03, for 100 iterations at a rate $\frac{1.5\times10^3}{\mathrm{CRPS}_0}$, and then for 40 iterations at a four times smaller rate to finalize the

convergence, a method called "step-decay schedule" for the learning rate (Ge et al. 2019). Finally, to go up to horizon 0.04, we repeat the procedure with 60 iterations at rate $\frac{1.5\times10^3}{\mathrm{CRPS}_0}$ and then 40 iterations at a rate four times smaller.

One thing that we expect from our algorithm is that the analog ensemble with the distance optimized to minimize the average CRPS would have a better representation of uncertainties. Indeed, when optimizing the RMSE, it is not necessary that the ensemble spread corresponds to the actual uncertainty of the forecast. In particular, other studies have reported RMSE-based optimization of machine-learning algorithms to give over-confident ensemble forecasts (Frion et al. 2024). In our case, we observe the opposite behaviour: the RMSE-based optimization is under-confident: the ensemble spread is larger than the actual uncertainty of the forecast. We interpret this fact as the consequence of a linear properties of the Lorenz System at this small forecast horizon and over the 200 selected analogs for this catalog size. In the case of linear dynamics, the average value is closer to the truth when including more members.

To evaluate the goodness of uncertainty quantification between the two types of optimization, we investigate first the prediction error, noted $z_{ana}(t+h) - z_{truth}(t+h)$ in Fig. 5.a, where the subscript "ana" refers to "analogs". This notation can be reconciled with the previous notations through the following identity:

$$z_{ana}(t+h) - z_{truth}(t+h) = y_j - y_i\,. \tag{18}$$

This variable would be zero ideally. In particular, the probabilities $p(j|i)$ should be highest when this variable is close to zero. To evaluate this, we compute the empirical histograms of this variable $y_j - y_i$, weighted by $p(j|i)$, for all values of i in the catalog and $j \in I(i)$, and comparing the results of the three distances (unoptimized, optimized for RMSE, optimized for CRPS) in Fig. 5.a. We can see that all distributions are centered, and the CRPS is giving the sharpest distribution, indicating a better uncertainty quantification.

Finally, in Fig. 5.b, we show the following variable:

$$F_{ana}(z_{truth}(t+h)) = \sum_{j\in I(i)} p(j|i)\mathbb{1}(y_j < y_i)\,, \tag{19}$$

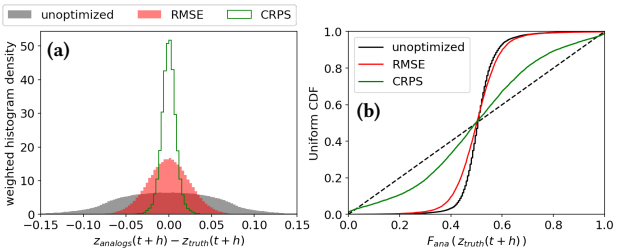

FIG. 5. Verifying statistics of analog forecasting ensembles for the case of the forecast of the z-variable of the Lorenz system at horizon h =0.04, computed on 10^4 points. (a) Empirical distribution of the difference between the analog forecast and the truth, for different distances : unoptimized (grey, full), optimized for RMSE minimization (red, full) and optimized for CRPS minimization (green, empty). Each value of $z_{ana}(t+h) - z_{truth}(t+h)$ is given a weight $p(\mathbf{x}_ana|\mathbf{x}_{truth}$ in the empirical density estimate. (b) Probability-probability plot (P-P plot) between the reference cumulative probability distribution of a uniform random variable (vertical axis) and the cumulative probability distribution of the analogue forecasts applied to the true value $z_{truth}(t+h)$. The dashed line indicates what a perfect forecast distribution would give.

which is the empirical cumulative distribution of the analog forecast applied to the true outcome value. Ideally, if the true outcome was drawn from the analog forecast distribution, the variable $F_{ana}(z_{truth}(t+h))$ would be uniformly distributed, which would mean that the true outcome value is exactly Q% of the time above the Q-percentile of the analog ensemble forecast distribution. In practice, the uncertainty estimation from the analog ensemble forecast distribution is not perfect, however one aim of our CRPS-oriented algorithm is to optimize the distance in order to improve uncertainty quantification from the analog ensemble. This is verified in Fig. 5.b which plots the quantiles of $F_{ana}(z_{truth}(t+h))$ versus the quantiles of a uniform distribution. In the ideal case, the empirical line would lie along the diagonal. In this figure it is clear that the empirical results from the RMSE-based optimization is closer to the diagonal line than the results with unoptimized isotropic distance, and that the results are even more satisfying with the CRPS-based optimization. However, further improvement is needed, as the figure still indicates an overdispersive ensemble: the uncertainty estimate from the analog ensemble is greater than the true uncertainty of the forecast.

4. Tropical cyclone intensity forecasting

a. Intensity forecasting

In this section we apply our algorithm to the study of the tropical cyclone, a meteorological system for which actual datasets, even at the global scale, are of a limited size. In such a context, we evaluate whether our method overcomes the typical dimensionality issues encountered in most tropical cyclone forecasting applications.

We use data from the International Best Track Archive for Climate Stewardship (IBTrACS, Knapp et al. 2010), a global compilation of best-track datasets from multiple international meteorological agencies which includes estimates of the storm location, intensity, and size in four geographical quadrants, on a six-hourly basis. IBTrACS suffers from spatio-temporal heterogeneities, especially concerning R_{max}, a critical parameter for the system dynamics (Avenas et al. 2024a), but it remains the most global dataset on tropical cyclones. In addition, statistical relationships have been recently developed to backup R_{max} estimates from better known parameters included in this global dataset (Avenas et al. 2023) and are used here.

For our experiment, we use only parts of the IBTrACS dataset. First, we restrict to the North-Atlantic ocean basin, to ensure a unified data treatment and definition from American agencies. Second, we use data from years 2003-2022, as we want to use a specific variable (R_{34}, see definition in appendix B) which is not included before 2003. Also, we consider solely tropical cyclones (TC) with a maximum wind speed V_{max} greater than 33m/s (see definition in appendix B) at least once in its lifetime. After this selection, we are left with 110 TCs re-sampled at 3h-time step (see appendix B).

The output variable to be estimated with analogs is defined as the difference in maximum sustained wind speed :

$$y := \Delta V_{max}(t,h) := V_{max}(t+h) - V_{max}(t)\,, \tag{20}$$

while the 15 input variables used to find and rate analogs are:

$$
\begin{aligned}
\mathbf{x} \coloneqq \Big\{ & V_{max}(t),\, R^{IBT}_{max}(t),\, R_{34}(t),\, f_{Cor}(t),\, u_{trans}(t), \\
& v_{trans}(t),\, R^{A23}_{max}(t),\, \frac{\mathrm{d}V_{max}}{\mathrm{d}t}(t),\, \frac{\mathrm{d}R^{IBT}_{max}}{\mathrm{d}t}(t), \\
& \frac{\mathrm{d}R_{34}}{\mathrm{d}t}(t),\, \frac{\mathrm{d}f_{Cor}}{\mathrm{d}t}(t),\, \frac{\mathrm{d}u_{trans}}{\mathrm{d}t}(t), \\
& \frac{\mathrm{d}v_{trans}}{\mathrm{d}t}(t),\, \frac{\mathrm{d}R^{A23}_{max}}{\mathrm{d}t}(t),\, T_{18}(t) \Big\} \ .
\end{aligned}
\tag{21}
$$

See appendix B for a description of these variables. As in the Lorenz-63 experiments, all variables are normalized by their standard-deviation. In the present work, we use our algorithm to minimize the CRPS of the analog ensemble forecast of $y = \Delta V_{max}(t, h)$, imposing the linear transform matrix $\mathbf{A}$ to be square-diagonal. Note that we make use of regularization terms only in the next section. Our algorithm is used here for the purpose of weighting the (normalized) input variables in $\mathbf{x}$, with 15 weighting coefficients corresponding to each diagonal element of $\mathbf{A}$.

Note that we do not wish to compete with state-of-the-art forecasting algorithms as this would require a whole dedicated study, in particular for the definition of the dataset and input variables. Rather, our interest is in evaluating the behaviour of our algorithm on a reanalysis and reduced dataset describing a real-life physical problem. Also, we wish to demonstrate the ability of our algorithm to fine-tune the weighting of variables, without prior physical knowledge, for the challenging task of tropical cyclone intensity forecasting (Emanuel and Zhang 2016; Cangialosi et al. 2020).

From the 110 TCs in our dataset, we use 73 TCs for training (2/3 of the dataset) and 37 TCs for test. The train/test splitting is random, and we repeat the experiment with 10 random splittings to assess the sensitivity of the results to the splitting choice. In the training phase (*i.e.*, the optimisation of $\mathbf{A}$), we use all the training dataset to evaluate the average CRPS and its gradient. We use a special type of leave-one-out methodology: to forecast $\Delta V_{max}(t, h)$ of a given TC, we search for analogs in other TCs, but we also allow the use of analogs from the same TC only with a minimum separation of ±3 days. This allows to raise the number of potential analogs and simplifies the algorithm structure. However, to assess the algorithm's performance on the test set, analogs are only searched within the training set, and therefore a given TC cannot be used as an analog of itself.

The good performances of the algorithm on the test set justify the reliability of this procedure (see below).

First, we run our algorithm without regularization (*i.e.* $\lambda = 0$, see Eq. 12), and for forecast horizons h = 12h, 24h, 36h, ... , 120h. For each horizon h, the transformation matrix is initialized with the identity matrix $\mathbf{A} = \mathbf{A}_{iso}$, and we use a constant learning rate of $\frac{10}{\text{CRPS}_0}$ where CRPS_0 is defined as previously as the average CRPS on the training set with $\mathbf{A} = \mathbf{A}_{iso}$. The algorithm is run for 50 iterations, which is enough to reach convergence. Note that the algorithm is run 10 times for each horizon as we use 10 random train/test splittings. We note CRPS_{con} the average CRPS obtained with $\mathbf{A} = \mathbf{A}_{con}$ the converged matrix, and we define the CRPS gain as:

$$\%\text{CRPS}_{gain} := 100\frac{\text{CRPS}_0 - \text{CRPS}_{con}}{\text{CRPS}_0} \ . \tag{22}$$

Note that this definition can be used for the average CRPS on the training set (using analogs from the training set to forecast TCs in the training set) and for the average CRPS on the test set (using analogs from the training set to forecast TCs in the test set). Ideally the two gains would be nearly identical, which would indicate that the weights optimized on the training set generalize well to the test set. This is confirmed in Fig. 6(b), where we show for each horizon h the median and percentiles of the CRPS gain on the 10 random train/set splittings. The CRPS gain is slightly higher on the training set, which is the sign of a slight overfitting, however the gains on the training and test sets are similar for every horizon and follow the same tendency of a growth of the gain with horizon. Gains are substantial, ranging from 7% to 20%, showing the interest of our methodology compared to the use of a brute-force unoptimized distance.

To assess the competitiveness of the analog methodology, we also compute the CRPS of two benchmarks: persistence and climatological forecasts. Here, we define persistence forecast as a deterministic forecast (*i.e.* a one-member ensemble) with $\Delta V_{max} = 0$, and its average CRPS is given by its mean absolute error. We define the climatological forecast as an ensemble forecast where all the elements of the dataset are used to build an ensemble forecast, and each element is given equal weights. The climatological forecast is therefore given by the whole empirical distribution of ΔV_{max}, and therefore depends on the forecast horizon h. These benchmarks are evaluated on the whole dataset (110 TCs), and compared to the (optimized and unoptimized) analog ensemble forecasts on the whole dataset (train and test) using the leave-one-out methodology described

earlier. The corresponding average CRPS are shown as a function of horizon in Fig. 6(a). Note that several black lines correspond to 10 different values of $\mathbf{A}_{con}$ for each horizon h, associated with the 10 random train/test splittings. The analog forecasts outperform the persistence and climatological forecasts, especially for large forecast horizons, which is also where the CRPS gain due to optimization is largest. The gain in CRPS for an horizon of 5 days thanks to our optimization is of ∼1.5m/s, which is close to the difference between the climatological forecast and the analog forecast with unoptimized distance (∼2m/s). Note that the climatological forecast can be viewed as an analog ensemble forecast, where the number of analogs equals the size of the catalog, and all distances are equal (*i.e.* $p(i|j)$ is flat, using notations from section 2a). This means that our algorithm allows for a gain in CRPS from a "naïve" (unoptimized) distance which is comparable to the gain obtained when passing from a "flat" distance (climatology) to a naïve (unoptimized) distance.

b. Variable selection

Then, we assess the ability of our algorithm to perform variable selection, using the regularization term introduced in Eq. (12). We do so for forecast horizon h = 1 day, and for regularization coefficients λ = 0, 0.001, 0.002, 0.003, ..., 0.015. For each value of λ, we take 10 random train/test splittings as previously. We run our algorithm on each training set with constant learning rate equal to $\frac{10}{\text{CRPS}_0}$ where CRPS_0 is defined as previously as the average CRPS on the training set with $\mathbf{A} = \mathbf{A}_{iso}$. The algorithm is run for 100 iterations this time, as the addition of the regularization term requires more iterations to converge.

To see which variable is selected by the algorithm as we vary λ, we show in Fig. 7 the average of each coefficient of $\mathbf{A}_{con}$ over the 10 randomly selected training sets for each value of λ. This shows that as λ grows, the smallest coefficients are drawn to zero, while the largest grow as a compensation. For these experiments, the algorithm has selected V_{max}, as well as the Coriolis frequency f_{Cor} although with a smaller weight. The meridional translation velocity is also selected as a statistically relevant feature. The value of $\frac{\mathrm{d}V_{max}}{\mathrm{d}t}$ is also statistically significant for the selection of analogs, which seems reasonable as the local 3h-velocity growth rate is likely to be informative of the short-term evolution.

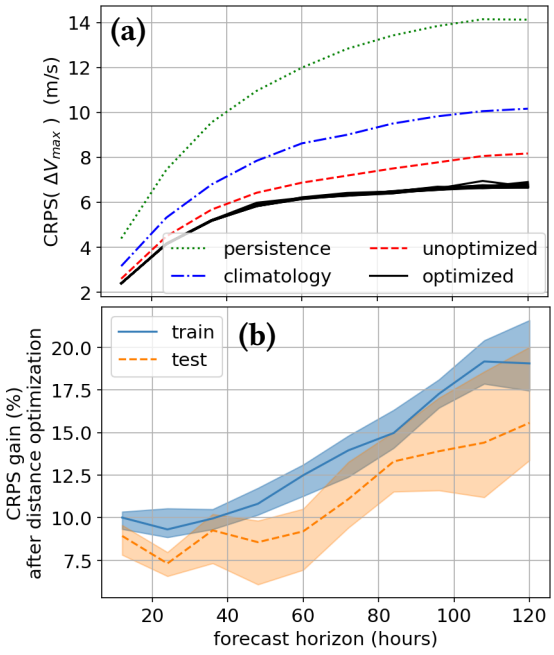

FIG. 6. (a) Average CRPS (meters per second) of the forecast of ΔV_{max} for different methods: persistence (deterministic forecast with $\Delta V_{max} = 0$), climatology (using the whole empirical distribution of ΔV_{max} as an ensemble), and unoptimized and optimized analog ensemble forecasts. (b) Gain in average CRPS of analog forecast after optimizing the distance, on the training set used to optimize the distance (blue, full line) and on the independent test set (orange, dashed line). The lines show the medians, while the shaded areas show the 25% and 75% percentiles.

However Fig. 7 shows an average over 10 realizations of the training set, but the results for each set can differ. Also, in practice, when doing variable selection, one is not interested in a particular value of λ but in a fixed number of variables. To take this practical point of view, we take all the values of converged $\mathbf{A}_{con}$ from this experiment, and rank them by the number of coefficients of $\mathbf{A}$ which are above 0.15, an arbitrary threshold which gives an idea of the number of variables selected by the algorithm. This allows to compute what is the average weight given to each variable when fixing the number of non-negligible coefficients (Fig. 8(a)), as well as the probability to pick a given variable when the number of non-negligible coefficients is fixed (Fig. 8(b)).

Although Fig. 8(a) mostly reproduces the behaviour of Fig. 7(a), some other features can be extracted from this figure. One first striking fact is that the algorithm never selects more than 11 variables out of 15, even when the regularization parameter is set to zero. This shows that

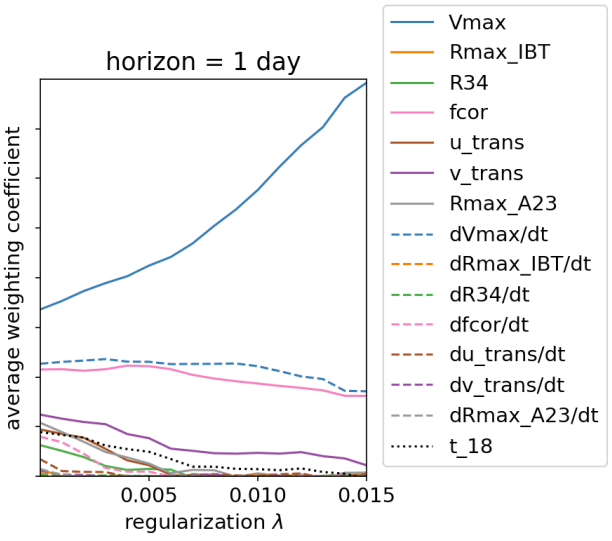

FIG. 7. Weighting coefficients after optimization for the h = 1day-forecast of ΔV_{max}, as a function of regularization parameter $\lambda > 0$. The coefficients are averaged over 10 random splittings of the dataset into training and test sets.

the algorithm is able to discard statistically irrelevant features without having to impose scarcity through regularization. Also, we witness some nontrivial behaviour, in particular the fact that $\frac{\mathrm{d}u_{trans}}{\mathrm{d}t}$ is always selected when using 10 variables, but never when using 11. This goes against the idea of the iterative algorithm of Alessandrini et al. (2018), described in the introduction. This other algorithm goes from right to left in our Fig. 8, starting by adjusting the coefficients of the most relevant variables used alone, and then adding new variables iteratively. Here, we show that a variable which is relevant when used with a certain number of variables may not be relevant when using a larger number of variables. From a methodological standpoint, this means that one should in principle select all relevant variables at once and not iteratively, one-by-one as proposed by Alessandrini et al. (2018).

Finally, note that the CRPS of the analog ensemble forecast with converged $\mathbf{A}_{con}$ is a decreasing function of λ: a higher regularization induces a larger error. For horizon h = 1 day, the decrease of CRPS($\mathbf{A}_{con}$) with λ is significant (at most a loss of 4.6% in CRPS$_{gain}$ for $\lambda = 0.015$, while the CRPS$_{gain}$ has values ~9%, not shown). We also do not witness a better generalization when applying regularization: the ratio between the CRPS evaluated on the test set and the one evaluated on the training set is not higher when raising λ. This shows that the number of variables selected by

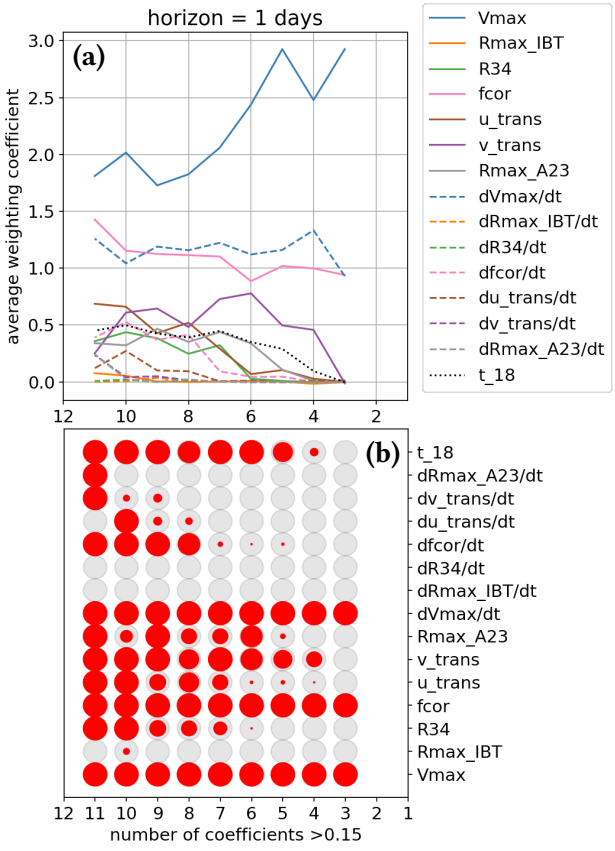

FIG. 8. (a) Average weighing coefficient for each input variable, when aggregating converged values of $\mathbf{A}_{con}$ for which the number of coefficients > 0.15 is fixed. (b) The horizontal axis is the same as in the top panel, while the vertical axis corresponds to the input variables. The radius of the red disks is proportional to the number of times that this variable is selected with a coefficient > 0.15 in all aggregated values of $\mathbf{A}_{con}$. When the red disk is as large as the grey disk, this indicates that the variable is always selected. In contrary, when there is no red disk in front of the grey disk, the variable is never selected.

the algorithm with $\lambda = 0$ is already sufficiently small for the analog ensemble forecast to generalize well, and imposing to use a lower number of variables is detrimental for forecast horizon $h = 1$ day.

Note that several input variables are redundant here. For instance, the rate of change of the Coriolis frequency $\frac{\mathrm{d}f_{Cor}}{\mathrm{d}t}$ can be expressed as a function of f_{Cor} and of the meridional translation velocity. Also, R_{max}^{A23} is statistically determined from V_{max}, f_{Cor} and R_{34}. The fact that these variables are still selected some times by the algorithm indicates that creating new variables out of existing ones for the definition of the distance may be useful in the case of analog forecasting, in particular if these variables are non-linear functions of the initial variables that are based on

previous physical or empirical analysis showing their relevance. This is typically the case of R_{max}^{A23} (see Eq. B3 in appendix B). Note also that this last variable is selected much more often than the R_{max}^{IBT} from the IBTrACS dataset, although R_{max}^{A23} is a statistical approximation for the true radius of maximum wind speed. This confirms both the limitations of this parameter in the IBTrACS database (Combot et al. 2020) and the utility of empirical approaches to overcome this issue when studying the tropical cyclone dynamics (see also the discussion in Avenas et al. 2023). However, these experiments on IBTrACS data show that defining relevant variables is not enough, and should be complemented by an approach such as the one proposed here to systematically, at least, tune the weights given to each variable and the overall scale of the distances used to rate analogs.

5. Conclusion and perspectives

We have shown algorithms originally developed in the field of "distance learning", which is a sub-field of machine-learning, can be adapted to allow for the optimization of the distance used in analog methods. To our knowledge, this is the first time that the gap between distance learning and analog methods is bridged. Our algorithm learns a linear transformation of the feature variables that is applied ahead of a classical Euclidean-distance-based analog ensemble methodology. This is equivalent to learning a "Mahalanobis-like" distance, but differs from using the Mahalanobis distance where the data's covariance matrix is used directly. Distance learning algorithms were initially designed for classification purposes, with little interest in quantify uncertainties. On the contrary, our algorithm is designed for continuous estimation purposes (regression), such as the forecasting or downscaling of scalar variables. Furthermore, our algorithm tunes the distance so that the analog ensemble gives an accurate estimation of uncertainty while staying as close as possible to the ground truth. This is done in particular through the use of the continuous-ranked probability score as a loss function.

Our tests of the algorithm on analog forecasts of the three-variable chaotic Lorenz System show a non-trivial dependency of the optimal distance with forecast horizon, as well as catalog size. For low-size datasets, we observe strong variations of the optimal distance with catalog size, followed by stabilization above a given threshold and eventually convergence. We also observe the growth of the scale of the optimal distance with forecast horizon, indicating that a more severe selection of analogs is needed for long-term forecasts. These examples show that the optimal distance

strongly depends on the system under study, the objective task, and the number of available data. Finally, we show that a CRPS-based optimization allows to have better uncertainty quantification from analog ensembles compared to RMSE-based optimization that were developed previously in distance learning algorithms. This demonstrates the benefit of our adapted algorithm for the case of analog methods in atmospheric and ocean science.

To investigate the behaviour of our algorithm on a real system, we use IBTrACS tropical cyclone data, and test the ability of our algorithm to weight input variables in the case of intensity forecasting. First, analog methods outperform simple methods such as persistence or climatological forecasts at all forecast horizons in terms of CRPS. Second, our algorithm allows for significant improvement with respect to a baseline of analog forecast where all input variables are given equal weights. Third, even without regularization our algorithm already removes some unrelevant input variables, allowing for a first dimension reduction. To further reduce the number of variables used we add a regularization term, allowing to reveal which variables contribute the most to the optimal definition of distance for analog ensemble forecasting of tropical cyclone intensity. This demonstrates that our algorithm can be used on small-size datasets, which is an interesting property for TC studies, and that our algorithm allows to perform dimension reduction which is a key requirement of analog methods.

We note here that extensions of this algorithm were tried but not retained. These include the definition of a state-dependant distance, where the matrix $\mathbf{A}$ is itself a (smooth) function of $\mathbf{x}$. This is one possible way of having a non-linear transformation, which is a generalization of our linear (constant) transformation $\mathbf{A}$. We have also tried to optimize several distance used at once, building several analog ensemble for each forecast, each with a different weight that is also optimized in the algorithm. Although feasible in practice, these extensions of our algorithm were computationally more intensive and did not yield remarkable improvements.

Our algorithm could still be modified in several ways to deal with existing problems in atmospheric and ocean science. First, note that the experiments conducted for this study were all performed on a personal laptop, but memory issues would arise in the case of both high number of features (>100) and large number of training samples, which could be the consequence of using gridded fields of geophysical variables as features. These memory issues would be due to the products $\mathbf{x}_{ij}\mathbf{x}_{ij}^T$ which is as large as the square of the number of features (unless $\mathbf{A}$ is diagonal). However,

there are memory-optimal ways to estimate such products, such as low-rank matrix approximations (Kumar et al. 2012). Also, in the experiments performed here we have used batch-gradient descent, which means that we use the whole training sample to compute the gradient (the whole sums over N in sections **2**.*b* and **2**.*c*), but other techniques such as stochastic gradient descent (Bottou 2012) or mini-batch gradient descent (Khirirat et al. 2017), which compute the gradient over subsets of the training set, would allow to diminish the memory requirements of the method. These techniques could also help escaping sub-optimal local minima, since we are facing a non-convex optimization problem. More generally, routines available from machine-learning libraries allow to compute gradients very efficiently, which already enabled us to perform experiments on feature vectors of size exceeding 10^3 on a personal laptop (not shown).

The case of extreme events could be tackled by using our algorithm for weighted-CRPS minimization and giving more weights to the large values. This has potential applications for statistical downscaling of extreme precipitation, for instance.

Finally, note that there are numerous distance learning algorithms that are different from the ones we have used and could also be modified to meet the requirements of analog methods in ocean and atmospheric science. In particular, some algorithms have the property of solving convex optimization problems, including for instance Globerson and Roweis (2005). We are currently working on such adaptations.

Acknowledgments. P. Platzer, B. Chapron and A. Avenas aknowledge the support from ERC project 856408-STUOD. A. Mouche acknowledges the support from ESA MAXSS (4000132954/20/I-NB) and ESA MPC-S1 projects (4000107360/12/I-LG).

Data availability statement. The data used and generated for this article are available upon request.

APPENDIX A

Generalization to weighted CRPS

An interesting feature of the CRPS is the possibility to weight the CRPS, and therefore give more importance to specific outcome (*e.g.*, extreme values). The weighted CRPS is defined as:

$$\mathrm{wCRPS}(F, y) = \int_{-\infty}^{\infty} \left[F(y') - \mathbb{1}(y' > y)\right]^2 w(y)\mathrm{d}y' \,, \tag{A1}$$

for any non-negative function $w(y)$, and can be further expressed as (Taillardat et al. 2023):

$$\mathrm{wCRPS}(F, y) = \mathbb{E}_F|W(Y) - W(y)| - \frac{1}{2}\mathbb{E}_F|W(Y) - W(Y')| \,, \tag{A2}$$

where $W(y) = \int_{-\infty}^{y} w(y')\mathrm{d}y'$ is any primitive of w. Using our notations, rewriting $W_{ij} := |\int_{y_i}^{y_j} w(y')\mathrm{d}y'|$, noting $\mathrm{wMAE}_i := \sum_{j\in I(i)} p(j|i)W_{ji}$ and $\mathrm{wMAD}_i := \sum_{j,k\in I(i)} p(j|i)p(k|i)W_{jk}$, we find that such a weighted CRPS has gradient:

$$\frac{\partial\overline{\mathrm{wCRPS}}}{\partial\mathbf{A}} = \mathbf{A}\frac{1}{N}\sum_{i=1}^{N}\sum_{j\in I(i)} p(j|i)\left\{\mathrm{wMAE}_i - \mathrm{wMAD}_i - \left(W_{ji} - \sum_{k\in I(i)} p(k|i)W_{kj}\right)\right\}\mathbf{x}_{ij}\mathbf{x}_{ij}^T \,. \tag{A3}$$

APPENDIX B

Variables used for tropical cyclone forecasting

Variables used in this paper for the tropical cyclone forecasting experiment come directly or indirectly from the IBTrACS database. IBTrACS is a compilation of the best-track data prepared by the different Regional Specialized Meteorological Centers (RSMCs) and Tropical Cyclone Warning Centers (TCWCs). Based on their area of responsibility, these regional agencies provide analyses of the TC location, intensity and structure on a regular time basis using the available data.

IBTrACS variables used in this study are the following:

- V_{max}, the "maximum sustained wind speed";
- R_{max}^{IBT} (we use the ^{IBT}-superscript to distinguish from R_{max}^{A23} introduced below) is the radius of maximum sustained wind speed, defined as the distance between the TC center and the position at which V_{max} is measured;
- R_{34}, the radius at which the velocity reaches 34 knots (1 kt $\approx$ 0.51 m/s) in four geographical quadrants (NE, SE, SW, and NW);
- lat, the latitude of the TC center;
- $storm\_speed$, the storm translation speed;
- $storm\_dir$, the storm translation direction.

Because of varying definitions of the maximum sustained wind speed across the different agencies, we selected only USA agencies (*i.e* National Hurricane Center, Joint Typhoon Warning Center, and Central Pacific Hurricane Center) which all provide the 1-minute maximum sustained wind speed.

Furthermore, to focus on the strongest storms, and to ensure well-defined R_{34} values, we removed all storms with a lifetime maximum intensity lower than 17.5 m/s. We also considered storms for which R_{max}^{IBT} was defined for at least 72 consecutive hours. Lastly, we cropped all storm time series to select the part of each events for which V_{max} was comprised between 17.5 m/s and the lifetime maximum intensity, to investigate the intensification period.

In IBTrACS, some storm tracks are given on a six-hourly basis, while others are interpolated and thus given on a three-hourly basis. After applying the procedure mentioned above, the 111 remaining storm tracks were all given on a three-hourly basis, except one, that was removed for consistency.

Then, the selected IBTrACS parameters have been directly used or transformed into other variables more relevant for the present study. The transformed variables include:

- R_{34}, whose nonzero values were averaged over the four geographical quadrants;
- f_{Cor}, the Coriolis frequency, defined as $f_{Cor} = 2\Omega\sin(lat)$, where $\Omega = 7.292x10^{-5}\ s^{-1}$ is the Earth angular velocity and lat is the latitude of the TC center;

- u_{trans} and v_{trans}, the TC translation speed in the zonal and meridional directions, computed with $storm\_speed$ and $storm\_dir$;

- $T_{18}(t)$, the number of hours after which V_{max} has reached 17.5 m/s. By definition, for each element of our dataset, $T_{18}(t) > 0$.

- R_{max}^{A23}, the radius of maximum wind speed estimated using the procedure described in Avenas et al. (2023) and detailed below.

The procedure to estimate R_{max}^{A23} can be summarized in three steps. An estimate of the TC maximum sustained wind speed that would correspond to an azimuthal average of the wind field, is first performed, using

$$V_{max,1D} = 0.6967 V_{max} + 6.1992. \tag{B1}$$

Second, the absolute angular momentum that an air parcel loses between R_{34} and R_{max} is estimated with the statistical relationship

$$\frac{M_{max,1D}}{M_{34}} = 0.531 \exp\{-0.00214(V_{max,1D} - 17.5ms^{-1}) - 0.00314(V_{max,1D} - 17.5ms^{-1})(\frac{1}{2} f_{Cor} R_{34})\}, \tag{B2}$$

where M_{34} is defined as $M_{34} = R_{34} * 17.5m/s + \frac{1}{2} f_{Cor} R_{34}^2$. Lastly, R_{max}^{A23} is estimated using the absolute angular momentum definition

$$R_{max}^{A23} = \frac{V_{max,1D}}{f_{Cor}} \left(\sqrt{1 + \frac{2 f M_{max,1D}}{V_{max,1D}^2}} - 1 \right). \tag{B3}$$

References

Alessandrini, S., L. Delle Monache, C. M. Rozoff, and W. E. Lewis, 2018: Probabilistic prediction of tropical cyclone intensity with an analog ensemble. *Monthly Weather Review*, **146 (6)**, 1723–1744.

Alexander, R., and D. Giannakis, 2020: Operator-theoretic framework for forecasting nonlinear time series with kernel analog techniques. *Physica D: Nonlinear Phenomena*, **409**, 132 520.

Avenas, A., B. Chapron, A. Mouche, P. Platzer, and L. Vinour, 2024a: Revealing short-term dynamics of tropical cyclone wind speeds from satellite synthetic aperture radar. *Scientific Reports*, **14 (1)**, 12 808.

Avenas, A., A. Mouche, J. Knaff, X. Carton, and B. Chapron, 2024b: On the tropical cyclone integrated kinetic energy balance. *Geophysical Research Letters*, **51 (16)**, e2024GL108 327.

Avenas, A., A. Mouche, P. Tandeo, J.-F. Piolle, D. Chavas, R. Fablet, J. Knaff, and B. Chapron, 2023: Reexamining the estimation of tropical cyclone radius of maximum wind from outer size with an extensive synthetic aperture radar dataset. *Monthly Weather Review*, **151 (12)**, 3169–3189.

Bellet, A., A. Habrard, and M. Sebban, 2022: *Metric learning*. Springer Nature.

Benestad, R. E., 2010: Downscaling precipitation extremes: Correction of analog models through pdf predictions. *Theoretical and Applied Climatology*, **100**, 1–21.

Bessafi, M., A. Lasserre-Bigorry, C. Neumann, F. Pignolet-Tardan, D. Payet, and M. Lee-Ching-Ken, 2002: Statistical prediction of tropical cyclone motion: An analog–cliper approach. *Weather and forecasting*, **17 (4)**, 821–831.

Bonnardot, F., H. Quetelard, G. Jumaux, M.-D. Leroux, and M. Bessafi, 2019: Probabilistic forecasts of tropical cyclone tracks and intensities in the southwest indian ocean basin. *Quarterly Journal of the Royal Meteorological Society*, **145 (719)**, 675–686.

Bottou, L., 2012: Stochastic gradient descent tricks. *Neural Networks: Tricks of the Trade: Second Edition*, Springer, 421–436.

Butcher, J. C., 1996: A history of runge-kutta methods. *Applied numerical mathematics*, **20 (3)**, 247–260.

Cangialosi, J. P., E. Blake, M. DeMaria, A. Penny, A. Latto, E. Rappaport, and V. Tallapragada, 2020: Recent progress in tropical cyclone intensity forecasting at the national hurricane center. *Weather and Forecasting*, **35 (5)**, 1913–1922.

Chen, P., H. Yu, B. Brown, G. Chen, and R. Wan, 2016: A probabilistic climatology-based analogue intensity forecast scheme for tropical cyclones. *Quarterly Journal of the Royal Meteorological Society*, **142 (699)**, 2386–2397.

Combot, C., A. Mouche, J. Knaff, Y. Zhao, Y. Zhao, L. Vinour, Y. Quilfen, and B. Chapron, 2020: Extensive high-resolution synthetic aperture radar (sar) data analysis of tropical cyclones: Comparisons with sfmr flights and best track. *Monthly Weather Review*, **148 (11)**, 4545 – 4563, https://doi.org/10.1175/MWR-D-20-0005.1.

Delle Monache, L., F. A. Eckel, D. L. Rife, B. Nagarajan, and K. Searight, 2013: Probabilistic weather prediction with an analog ensemble. *Monthly Weather Review*, **141 (10)**, 3498–3516.

Elliott, R., 1943: Studies of persistent regularities in weather phenomena. *Synoptic Weather Types of North America*.

Elsberry, R. L., and H.-C. Tsai, 2014: Situation-dependent intensity skill metric and intensity spread guidance for western north pacific tropical cyclones. *Asia-Pacific Journal of Atmospheric Sciences*, **50**, 297–306.

Emanuel, K., and F. Zhang, 2016: On the predictability and error sources of tropical cyclone intensity forecasts. *Journal of the Atmospheric Sciences*, **73 (9)**, 3739–3747.

Fetanat, G., A. Homaifar, and K. R. Knapp, 2013: Objective tropical cyclone intensity estimation using analogs of spatial features in satellite data. *Weather and forecasting*, **28 (6)**, 1446–1459.

Fraedrich, K., C. C. Raible, and F. Sielmann, 2003: Analog ensemble forecasts of tropical cyclone tracks in the australian region. *Weather and forecasting*, **18 (1)**, 3–11.

Fraedrich, K., and B. Rückert, 1998: Metric adaption for analog forecasting. *Physica A: Statistical Mechanics and its Applications*, **253 (1-4)**, 379–393.

Frion, A., L. Drumetz, G. Tochon, M. D. Mura, and A. A. E. Bey, 2024: Koopman ensembles for probabilistic time series forecasting. *arXiv preprint arXiv:2403.06757*.

Ge, R., S. M. Kakade, R. Kidambi, and P. Netrapalli, 2019: The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares. *Advances in neural information processing systems*, **32**.

Globerson, A., and S. Roweis, 2005: Metric learning by collapsing classes. *Advances in neural information processing systems*, **18**.

Goldberger, J., G. E. Hinton, S. Roweis, and R. R. Salakhutdinov, 2004: Neighbourhood components analysis. *Advances in neural information processing systems*, **17**.

Hastie, T., 2009: The elements of statistical learning: data mining, inference, and prediction. Springer.

Held, M., P. Wolfe, and H. P. Crowder, 1974: Validation of subgradient optimization. *Mathematical programming*, **6**, 62–88.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, **15 (5)**, 559–570.

Hersbach, H., and Coauthors, 2020: The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, **146 (730)**, 1999–2049.

Hoerl, A. E., and R. W. Kennard, 1970: Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12 (1)**, 55–67.

Horton, P., M. Jaboyedoff, and C. Obled, 2017: Global optimization of an analog method by means of genetic algorithms. *Monthly Weather Review*, **145 (4)**, 1275–1294.

Jackson, C., T. Ruff, J. Knaff, A. Mouche, and C. Sampson, 2021: Chasing cyclones from space. *Eos*, **102**, –, https://doi.org/10.1029/2021EO159148.

Jézéquel, A., P. Yiou, and S. Radanovics, 2018: Role of circulation in european heatwaves using flow analogues. *Climate dynamics*, **50 (3-4)**, 1145–1159.

Khirirat, S., H. R. Feyzmahdavian, and M. Johansson, 2017: Mini-batch gradient descent: Faster convergence under data sparsity. *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, IEEE, 2880–2887.

Kingma, D. P., and J. Ba, 2014: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Knapp, K. R., M. C. Kruk, D. H. Levinson, H. J. Diamond, and C. J. Neumann, 2010: The international best track archive for climate stewardship (ibtracs) unifying tropical cyclone data. *Bulletin of the American Meteorological Society*, **91 (3)**, 363–376.

Krick, I. P., 1942: *A Dynamical Theory of the Atmospheric Circulation and Its Use in Weather Forecasting...: Studies of Persistent Regularities in Weather Phenomena*. California Institute of Technology.

Kumar, S., M. Mohri, and A. Talwalkar, 2012: Sampling methods for the nyström method. *The Journal of Machine Learning Research*, **13 (1)**, 981–1006.

Langmack, H., K. Fraedrich, and F. Sielmann, 2012: Tropical cyclone track analog ensemble forecasting in the extended australian basin: Nwp combinations. *Quarterly Journal of the Royal Meteorological Society*, **138 (668)**, 1828–1838.

Le Bras, P., F. Sévellec, P. Tandeo, J. Ruiz, and P. Ailliot, 2024: Selecting and weighting dynamical models using data-driven approaches. *Nonlinear Processes in Geophysics*, **31 (3)**, 303–317.

Lewis, W. E., T. L. Olander, C. S. Velden, C. Rozoff, and S. Alessandrini, 2021: Analog ensemble methods for improving satellite-based intensity estimates of tropical cyclones. *Atmosphere*, **12 (7)**, 830.

Lguensat, R., P. Tandeo, P. Ailliot, M. Pulido, and R. Fablet, 2017: The analog data assimilation. *Monthly Weather Review*, **145 (10)**, 4093–4107.

Lorenz, E. N., 1956: *Empirical orthogonal functions and statistical weather prediction*, Vol. 1. Massachusetts Institute of Technology, Department of Meteorology Cambridge.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *Journal of atmospheric sciences*, **20 (2)**, 130–141.

Lorenz, E. N., 1969: Atmospheric predictability as revealed by naturally occurring analogues. *Journal of Atmospheric Sciences*, **26 (4)**, 636–646.

Lucarini, V., and Coauthors, 2016: *Extremes and recurrence in dynamical systems*. John Wiley & Sons.

Matulla, C., X. Zhang, X. Wang, J. Wang, E. Zorita, S. Wagner, and H. Von Storch, 2008: Influence of similarity measures on the performance of the analog method for downscaling daily precipitation. *Climate Dynamics*, **30**, 133–144.

McDermott, P. L., and C. K. Wikle, 2016: A model-based approach for analog spatio-temporal dynamic forecasting. *Environmetrics*, **27 (2)**, 70–82.

McLachlan, G. J., 1999: Mahalanobis distance. *Resonance*, **4 (6)**, 20–26.

Neumann, C. J., and J. R. Hope, 1972: Performance analysis of the hurran tropical cyclone forecast system. *Monthly Weather Review*, **100 (4)**, 245–255.

Nicolis, C., 1998: Atmospheric analogs and recurrence time statistics: Toward a dynamical formulation. *Journal of the atmospheric sciences*, **55 (3)**, 465–475.

Peterson, L. E., 2009: K-nearest neighbor. *Scholarpedia*, **4 (2)**, 1883.

Platzer, P., P. Yiou, P. Naveau, J.-F. Filipot, M. Thiébaut, and P. Tandeo, 2021a: Probability distributions for analog-to-target distances. *Journal of the Atmospheric Sciences*, **78 (10)**, 3317–3335.

Platzer, P., P. Yiou, P. Naveau, P. Tandeo, J.-F. Filipot, P. Ailliot, and Y. Zhen, 2021b: Using local dynamics to explain analog forecasting of chaotic systems. *Journal of the Atmospheric Sciences*, **78 (7)**, 2117–2133.

Sauer, T., J. A. Yorke, and M. Casdagli, 1991: Embedology. *Journal of statistical Physics*, **65**, 579–616.

Taillardat, M., A.-L. Fougères, P. Naveau, and R. De Fondeville, 2023: Evaluating probabilistic forecasts of extremes using continuous ranked probability score distributions. *International Journal of Forecasting*, **39 (3)**, 1448–1459.

Tandeo, P., and Coauthors, 2015: Combining analog method and ensemble data assimilation: application to the lorenz-63 chaotic system. *Machine Learning and Data Mining Approaches to Climate Science: proceedings of the 4th International Workshop on Climate Informatics*, Springer, 3–12.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of cmip5 and the experiment design. *Bulletin of the American meteorological Society*, **93 (4)**, 485–498.

Tibshirani, R., 1996: Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **58 (1)**, 267–288.

Toth, Z., 1991: Intercomparison of circulation similarity measures. *Monthly weather review*, **119 (1)**, 55–64.

Tsai, H.-C., and R. L. Elsberry, 2014: Applications of situation-dependent intensity and intensity spread predictions based on a weighted analog technique. *Asia-Pacific Journal of Atmospheric Sciences*, **50**, 507–518.

Tsai, H.-C., and R. L. Elsberry, 2019: Combined three-stage 7-day weighted analog intensity prediction technique for western north pacific tropical cyclones: Demonstration of optimum performance. *Weather and Forecasting*, **34 (6)**, 1979–1998.

Van den Dool, H., 1994: Searching for analogues, how long must we wait? *Tellus A*, **46 (3)**, 314–324.

Weickmann, L., 1924: Wellen im luftmeer. *Treatise in Math.-Phys. of the Saxon Academy of Science*, **39 (2)**.

Weinberger, K. Q., and G. Tesauro, 2007: Metric learning for kernel regression. *Artificial intelligence and statistics*, PMLR, 612–619.

Yang, W., K. Wang, and W. Zuo, 2012: Fast neighborhood component analysis. *Neurocomputing*, **83**, 31–37.

Yin, P., E. Esser, and J. Xin, 2014: Ratio and difference of l_1 and l_2 norms and sparse representation with coherent dictionaries. *Communications in Information and Systems*, **14 (2)**, 87–109.

Yiou, P., 2014: Anawege: a weather generator based on analogues of atmospheric circulation. *Geoscientific Model Development*, **7 (2)**, 531–543.

Yiou, P., M. Boichu, R. Vautard, M. Vrac, S. Jourdain, E. Garnier, F. Fluteau, and L. Menut, 2014: Ensemble meteorological reconstruction using circulation analogues of 1781–1785. *Climate of the Past*, **10 (2)**, 797–809.

Yiou, P., and A. Jézéquel, 2020: Simulation of extreme heat waves with empirical importance sampling. *Geoscientific Model Development*, **13 (2)**, 763–781.

Zhao, Z., and D. Giannakis, 2016: Analog forecasting with dynamics-adapted kernels. *Nonlinearity*, **29 (9)**, 2888.

Zhen, Y., P. Tandeo, S. Leroux, S. Metref, T. Penduff, and J. Le Sommer, 2020: An adaptive optimal interpolation based on analog forecasting: application to ssh in the gulf of mexico. *Journal of Atmospheric and Oceanic Technology*, **37 (9)**, 1697–1711.

Zorita, E., and H. Von Storch, 1999: The analog method as a simple statistical downscaling technique: Comparison with more complicated methods. *Journal of climate*, **12 (8)**, 2474–2489.