

A barycenter-based approach for the multi-model ensembling of subseasonal forecasts

Camille Le Coz¹, Alexis Tantet¹, Rémi Flamary², and Riwal Plougonven¹

¹LMD/IPSL, École Polytechnique, Institut Polytechnique de Paris, ENS, PSL Research University, Sorbonne Université, CNRS, Palaiseau France

²CMAF, Ecole Polytechnique, Institut Polytechnique de Paris, CNRS, Palaiseau, France

Correspondence: Camille Le Coz (camille.le-coz@lmd.ipsl.fr)

Abstract. Ensemble forecasts and their combination are examined from the perspective of probability spaces. By treating ensemble forecasts as discrete probability distributions, multi-model ensemble (MME) forecasts are reformulated as barycenters of these distributions. We consider two barycenters, each defined with respect to a different distance metric: the L_2 barycenter, which corresponds to the traditional pooling method, and the Wasserstein barycenter, which better preserves certain geometric

5 properties of the input ensemble distributions.

As a proof of concept, we apply the L_2 and Wasserstein barycenters to the combination of four models from the Subseasonal to Seasonal (S2S) prediction project database. Their performance is evaluated for the prediction of weekly 2 m temperature, 10 m wind speed, and 500 hPa geopotential height over European winters. By construction, both barycenter-based MMEs have the same ensemble mean but differ in their representation of forecast uncertainty. In particular, the L_2 barycenter has a larger

10 ensemble spread, making it more prone to underconfidence. Although both methods perform similarly on average in terms of the Continuous Ranked Probability Score (CRPS), the Wasserstein barycenter performs better more frequently.

1 Introduction

Multi-model ensemble methods (MME)

Multi-model ensemble (MME) methods have been shown to improve forecast skill for different time scales, from short-range

15 (Heizenreder et al., 2006; Casanova and Ahrens, 2009) and medium-range (Hamill, 2012; Hagedorn et al., 2012) to seasonal forecasting (Palmer et al., 2004; Alessandri et al., 2011; Kirtman et al., 2014). The added value of MME over single-model ensemble (SME) forecasts has been attributed to several factors. First, there is in general no “best” SME (Hagedorn et al., 2005), the relative performance of the SMEs varies depending on the target considered (i.e. region and variable of interest, metrics, etc.). The MME can take advantage of the complementary skills of the SMEs, thereby performing better on average.

20 Second, Hagedorn et al. (2005) identified error cancellation and non-linearity of the metric as the main reasons for the MME’s performance exceeding the average performance of SMEs. Third, MMEs allow us to explore a new dimension of uncertainty that remains unexplored by SMEs. Combining different models enables us to account for uncertainty due to *model formulation*: by construction, an SME takes into account uncertainties in model initialization and can introduce variations in parameteriza-

tions to sample some of the uncertainty due to parameterization choices, but it cannot account for uncertainty due to model
25 formulation. In addition, Weigel et al. (2008) showed that MMEs can improve predictive skill if and only if the SMEs “fail to
capture the full amount of forecast uncertainty”. Finally, MMEs benefit from their larger ensemble size, which often implies a
better sampling of the underlying probability distribution.

MME for subseasonal to seasonal (S2S) forecasts

Subseasonal to seasonal forecasting bridges the gap between weather (medium-range) and seasonal forecasting. It corresponds
30 to the time range from two weeks up to two months. Predictions on this time scale are the focus of the Subseasonal to Seasonal
(S2S) prediction project, whose objectives are to improve forecast skills and promote their use (Vitart et al., 2017). As part of
this project, a database containing S2S forecasts from thirteen (originally eleven) operational centers has been made available
to the research community. One of the main research questions the S2S project aims to address is: “What is the benefit of a
multi-model forecast for subseasonal to seasonal prediction and how can it be constructed and implemented?” (Vitart et al.,
35 2017).

Several studies have investigated the potential benefits of MMEs for S2S forecasts (Vigaud et al., 2017, 2020; Specq et al.,
2020; Wang et al., 2020; Materia et al., 2020; Zheng et al., 2019; Pegion et al., 2019). These studies use different MME
methods, variables, and evaluation criteria, but they all concur that MMEs generally perform as well as, or better than, SMEs.
Moreover, Vigaud et al. (2017) and Specq et al. (2020) suggest that MMEs improve not only the skill but also the reliability
40 of probabilistic forecasts. However, studies using the pooling method (i.e. the concatenation of ensembles) point out that the
improved performance of MMEs is partly related to their larger number of members (Zheng et al., 2019; Specq et al., 2020;
Wang et al., 2020). Among other studies, Karpechko et al. (2018) investigates a specific Sudden Stratospheric Warming event
with a MME, while Ferrone et al. (2017) evaluates several MME methods, but neither compares the skill of MMEs to that of
SMEs. Thus, while the potential benefits of MMEs at the subseasonal scale have been clearly established, the question of how
45 best to combine ensemble forecasts from different models has received relatively little attention. The present study focuses on
this question.

The most direct and commonly used method for multi-model combination is the “pooling method”, which simply concate-
nates the ensemble members from the different models. The members of the resulting multi-model ensemble can either be
equally weighted or be assigned different weights based on the model skills (e.g. Weigel et al. (2008) for seasonal scale and
50 Wanders and Wood (2016) for subseasonal scale). In the aforementioned studies, all use some variation of this method except
Vigaud et al. (2017, 2020) and Ferrone et al. (2017). The latter focuses on the prediction of terciles, and therefore couple
multi-model combination with other post-processing methods (e.g. model output statistics method such as extended logistic re-
gression) to directly predict the terciles. Also focusing on quantiles, Gonzalez et al. (2021) uses a sequential learning algorithm
to linearly combine predictors (from the SMEs, as well as from climatology and persistence), with weights updated at each step
55 based on previous performance. Other methods for weighted-MME have been explored, but have not yet been applied at the
subseasonal scale. At the seasonal scale, Rajagopalan et al. (2002) developed a Bayesian methodology to combine ensemble
forecasts for categorical predictands (also used by Robertson et al. (2004) and Weigel et al. (2008) at the seasonal scale to

obtain terciles of the MME). Other approaches, such as the Ensemble Model Output Statistics (EMOS) and Bayesian Model Averaging (BMA), adopt a probabilistic perspective and aim to construct one probability density function (PDF) from the different ensemble forecasts. The EMOS method assumes a specific parametric form for this predictive PDF, the parameters being obtained from the input ensembles by optimizing with respect to a chosen score (e.g. the CRPS) over a training period (Gneiting et al., 2005). In contrast, the BMA assumes separate parametric forms for the PDFs of the input models (based on the ensemble forecasts). The predictive PDF is then a weighted mixture of these input PDFs, with weights given by the posterior probabilities of the input models (Raftery et al., 2005). The following framework of barycenter encompasses the BMA as will be discussed later.

MME as barycenters of discrete distributions

In this study, we propose to revisit the combination of multiple ensemble models from a different perspective. We consider each ensemble forecast as a discrete probability distribution and reformulate the multi-model ensemble as the barycenter of these distributions. The barycenter then serves as a tool to merge several ensemble distributions into a single transformed ensemble distribution. We show that, in this framework, the pooled MME is actually the (weighted) barycenter with respect to the L_2 distance. Since we work with distributions instead of a collection of ensemble members, the notion of barycenter can be extended to other metrics in the space of probability distributions. In particular, a natural distance in this space is the Wasserstein distance that stems from optimal transport theory (Villani, 2003). The Wasserstein distance is defined as the cost of the optimal transport between two distributions.

Optimal transport and the Wasserstein distance have been used in diverse applications in climate and weather. They have been used to measure the response of climate attractors to different forcings (Robin et al., 2017), to evaluate the performance of different climate models (Vissio et al., 2020), or to assess different parameterizations (Vissio and Lucarini, 2018). Moreover, Papayiannis et al. (2018) used Wasserstein barycenters to point-downscale wind speed from an atmospheric model. Robin et al. (2019) developed a multivariate bias correction method based on optimal transport, while Ning et al. (2014) used it within the framework of data assimilation to deal with structural error in forecasts.

Here, we use the Wasserstein barycenter, or more precisely its Gaussian approximation, as a tool to build multi-model ensembles and compare it to the more traditional pooling method. The Wasserstein distance has some interesting properties in probability distribution space. While the L_2 distance measures the “vertical” distance between distributions, the Wasserstein distance is based on the “horizontal” displacement between them. This allows the Wasserstein barycenter to retain some geometric properties of the input distributions (such as normality, for example, see Backhoff-Veraguas, Julio et al. (2022)). We investigate the impact of changing the metric (from L_2 to Wasserstein) on the MME performance by applying both barycenters to the combination of four models from the S2S database.

The paper is organized as follows. The use of the two barycenters as MME methods is presented in Section 2. We first link the pooling method to the L_2 -barycenter (2.2.1) before introducing the Wasserstein distance and its barycenter (2.2.2). We apply these methods to the combination of four models from the S2S database (namely, ECMWF, NCEP, ECCC, and KMA; Vitart et al., 2017). The case study is described in Section 3, including the datasets and evaluation metrics. The skill of the two

MMEs and four SMEs is evaluated and compared in Section 4. The results are discussed in Section 5, and the main conclusions are presented in Section 6.

2 Multi-model ensemble methods

95 2.1 Ensemble forecasts as discrete probability distributions

At the S2S time scale, it is necessary to move from a deterministic to a probabilistic approach using ensemble forecasting (e.g. Kalnay (2003)). The ensemble is typically generated by perturbing the initial conditions and running the model for each of these perturbed states. The set of perturbed initial conditions represents the initial uncertainty associated with possible errors in the initial state of the atmosphere. This initial uncertainty is then propagated forward in time by the model. Thus, an ensemble
100 forecast is a set of N perturbed forecasts representing the evolution in time of the probability density of atmospheric variables according to the model formulation.

An ensemble forecast aims to sample the probability distribution of the forecasted variable conditioned on the initial distribution. Now, this can also be considered as a discrete probability distribution μ on the state space $\mathbb{R}^{n_t}$ such that

$$\mu = \sum_{i=1}^N a_i \delta_{\mathbf{x}_i}$$

where N is the number of members in the ensemble, $\mathbf{x}_i \in \mathbb{R}^{n_t}$ is the position of the Dirac corresponding to the time series of the i -th member, n_t is the number of lead times, and a_i is the weight of the i -th member (such that for all $i \leq N, a_i \geq 0$ and $\sum_{i=1}^N a_i = 1$). Thus, here, $\mathbf{x}_i$ does not represent an instantaneous state, but a full time series. That is, μ is a multivariate
105 distribution where the n_t variables correspond to the forecasted values at different lead-times. In a standard ensemble forecast, all the members are equiprobable, so they have equal weights $a_i = \frac{1}{N}$. In the remainder of this section, we will look at ensemble forecast time-series from the perspective of discrete distributions.

2.2 Barycenters for multi-model combination

The goal of multi-model ensemble methods can be rephrased as combining the complementary information from these distributions to obtain a new discrete distribution that better represents the true probability distribution function of the forecasted
110 variable. A way to summarize a collection of distributions $(\mu_1, \mu_2, \dots, \mu_K)$ is to compute their barycenter. The barycenter is found by solving the following minimization problem

$$\arg \min_{\mu} \sum_{k=1}^K \lambda_k \cdot d(\mu_k, \mu)^2 \quad (1)$$

where λ_k represents the weights given to the distributions, and $d(\cdot, \cdot)$ is a distance between distributions. The barycenter, also
115 known as the Fréchet mean, is effectively the distribution that best represents the input distributions with respect to a criterion given by the chosen distance d .

2.2.1 L_2 barycenter

The L_2 distance between two distributions μ_1 and μ_2 is given by $\|\mu_1 - \mu_2\|_2 = \left(\int_{\mathbb{R}^{n_t}} (\mu_1(\mathbf{z}) - \mu_2(\mathbf{z}))^2 d^{n_t} \mathbf{z} \right)^{1/2}$. When using this distance in the barycenter equation (1), one can find the analytical formula for the L_2 barycenter of the distributions

120 $\mu_1, \dots, \mu_K$:

$$\mu_{L_2}^\lambda = \sum_{k=1}^K \lambda_k \mu_k \quad (2)$$

$$= \sum_{k=1}^K \lambda_k \sum_{i=1}^{N_k} a_{k,i} \delta_{\mathbf{x}_{k,i}} \quad (3)$$

where the distribution μ_k has N_k Dirac masses located at points $(\mathbf{x}_{k,1}, \dots, \mathbf{x}_{k,N_k}) \in \mathbb{R}^{n_t, N_k}$ with associated probability vector $\mathbf{a}_k = (a_{k,1}, \dots, a_{k,N_k}) = (\frac{1}{N_k}, \dots, \frac{1}{N_k})$ (for $k = 1, \dots, K$).

125 Equation (2) shows that the L_2 barycenter is a weighted average of distributions. This is similar to the BMA method, which can then be seen as a L_2 barycenter with specific weights derived from the Bayesian framework (in which the weights are the posterior probabilities of the input models, i.e., they represent the probability that the associated model is the best model). However, Equation (3) differs from BMA in the particular distributions being summed: an assumption on their shape is made in the BMA, while the ensemble distributions are directly used in the L_2 barycenter (Raftery et al., 2005).

130 For a generic choice of weights λ_k , one can see that the L_2 barycenter corresponds to the concatenation of the members of the K input ensembles with the rescaling of the associated member weights (see Eq. (3)). This is equivalent to “pooling” together the ensemble members from the different models. The pooling method is a simple and well-established MME method. It has been used at the seasonal scale (Hagedorn et al., 2005; Weigel et al., 2008; Robertson et al., 2004; Becker et al., 2014), at the decadal scale (Smith et al., 2013) and at the medium-range scale (Hamill, 2012; Hagedorn et al., 2012). More recently, it
 135 has also been applied to subseasonal ensemble forecasts by Karpechko et al. (2018); Pegion et al. (2019); Zheng et al. (2019); Specq et al. (2020); Materia et al. (2020). The pooling method is based on the assumption that all members from all ensembles are independent and identically distributed samples of the same distribution. Under this assumption, the members can be pooled to form an empirical distribution. However, this hypothesis is rarely verified in practice (Kioutsioukis and Galmarini, 2014; Knutti et al., 2017). If we assume instead that the input ensembles have different distributions because of model-specific errors
 140 (which we aim to sample), then interpreting the pooling method as a L_2 barycenter allows us to extend it to different distances. In the following, we introduce another distance, the Wasserstein distance, and its associated barycenter. From this point on, we consistently use L_2 barycenter to refer to pooling in order to facilitate comparison with this second barycenter-based MME method.

2.2.2 Wasserstein barycenter

145 The Wasserstein distance stems from optimal transport theory and can be seen as the cost of transportation between two distributions μ_1 and μ_2 . It can be defined on discrete distributions as:

$$W_2^2(\mu_1, \mu_2) = \min_{\mathbf{T} \in U(\mathbf{a}_1, \mathbf{a}_2)} \langle \mathbf{T}, \mathbf{C} \rangle = \min_{\mathbf{T} \in U(\mathbf{a}_1, \mathbf{a}_2)} \sum_{i,j} t_{i,j} \|\mathbf{x}_{1,i} - \mathbf{x}_{2,j}\|^2 \quad (4)$$

where $U(\mathbf{a}_1, \mathbf{a}_2) = \left\{ \mathbf{T} \in \mathbb{R}_+^{N_1, N_2} : \mathbf{T} \mathbf{1}_{N_2} = \mathbf{a}_1 \text{ and } \mathbf{T}^T \mathbf{1}_{N_1} = \mathbf{a}_2 \right\}$ is the set of all feasible transport matrices $\mathbf{T} = (t_{i,j})_{1 \leq i \leq N_1, 1 \leq j \leq N_2}$ between the probability vectors $\mathbf{a}_1$ and $\mathbf{a}_2$ associated with μ_1 and μ_2 respectively (with $\mathbf{1}_N$ standing for the all-ones vector of size N), and $\mathbf{C} \in \mathbb{R}^{N_1 \times N_2}$ is a distance matrix whose elements are the pairwise squared Euclidean distances between the elements of $\mathbf{x}_{1,:}$ and $\mathbf{x}_{2,:}$. Here, $(\langle \mathbf{T}, \mathbf{C} \rangle)^{1/2}$ is the cost associated with the transport $\mathbf{T}$ (with $\langle \cdot, \cdot \rangle$ being the Frobenius inner product between matrices). The elements $t_{i,j}$ of a transport matrix $\mathbf{T}$ describe the amount of mass going from $\mathbf{x}_{1,i}$ to $\mathbf{x}_{2,j}$, while $\|\mathbf{x}_{1,i} - \mathbf{x}_{2,j}\|^2$ is the cost of moving one unit of mass from $\mathbf{x}_{1,i}$ to $\mathbf{x}_{2,j}$. The minimization problem consists of searching for the optimal transport among all the feasible ones in $U(\mathbf{a}, \mathbf{b})$, i.e., the transport associated with the lowest cost denoted as the 2-Wasserstein distance. This is a short description of the Wasserstein distance for discrete distributions; for more information see Peyré and Cuturi (2020), Santambrogio (2015) or Villani (2003).

Contrary to the L_2 barycenter above, there is no general closed-form expression for the Wasserstein barycenter, and one must solve the minimization problem (1) to find it. However, computing the Wasserstein barycenter directly on discrete distributions causes an artificial decrease of the barycenter's variance, which leads to important under-dispersion of the multi-model ensembles. This variance shrinkage, illustrated in appendix A, is due to the limited sample size of these distributions with respect to the number of dimensions (corresponding here to the number of lead times, see Section 2.1).

In order to avoid the variance shrinkage and have a more efficient and better statistical estimation, we propose to estimate the barycenter using a Gaussian approximation of the data and then to “map” all forecast onto this barycenter using the Gaussian mapping. That is, a Gaussian assumption is made to estimate a Gaussian Wasserstein barycenter which provides a closed-form affine mapping from each distribution to the barycenter, and then this mapping is applied to the discrete distributions which aligns them on the barycenter but keeps their individual samples. This means that only the first- and second-order moments of the Gaussian are estimated from the data which attenuates the statistical difficulty of estimating the Wasserstein distance (Flamary et al., 2020). This approach was first used in Gnassounou et al. (2023) for data normalization of biomedical signals. Note that they additionally assumed the signal to be stationary and periodic to simplify the computations, however, these hypotheses are not appropriate to our dataset. The method is illustrated in Figure 1 and is made up of four steps:

1. Each discrete distribution μ_k is approximated by a multivariate normal distribution $\hat{\mu}_k$ of mean $\hat{\mathbf{m}}_k$ and covariance $\hat{\Sigma}_k$ (for $k = 1, \dots, K$), which are estimated by the unbiased sample mean and covariance respectively.

Note: in this study, the multivariate distribution does not represent the joint distribution between several meteorological variables, but the joint distribution between different time steps of the same meteorological variable (see Section 2.1).

- 175 2. The Wasserstein barycenter between the normal distributions is computed. It is also a multivariate normal distribution with mean $\hat{\mathbf{m}}_b = \sum_{k=1}^K \lambda_k \hat{\mathbf{m}}_k$ and covariance $\hat{\Sigma}_b$ that is estimated iteratively using the fixed-point iteration $\hat{\Sigma}_b = \sum_{k=1}^K \lambda_k \left(\hat{\Sigma}_b^{1/2} \hat{\Sigma}_k^{-1/2} \hat{\Sigma}_b^{1/2} \right)^{1/2}$ from Agueh and Carlier (2011).
- 180 3. The optimal transports from each input Gaussian distribution $\hat{\mu}_k$ to the Gaussian barycenter $\hat{\mu}_b$ are derived. They are given by the following affine mappings $T_k(\mathbf{x}) = \mathbf{A}_k(\mathbf{x} - \hat{\mathbf{m}}_k) + \hat{\mathbf{m}}_b$ with $\mathbf{A}_k = \hat{\Sigma}_k^{-1/2} \left(\hat{\Sigma}_k^{1/2} \hat{\Sigma}_b \hat{\Sigma}_k^{1/2} \right)^{1/2} \hat{\Sigma}_k^{-1/2}$ (assuming that the $\hat{\Sigma}_k$ are invertible, see below).
4. The mappings are applied to the corresponding discrete distributions (that is the mapping T_k is applied to each state member $\mathbf{x}_{k,i}$ of μ_k), and then the adapted distributions are pooled together (as for the L_2 barycenter in eq. (3)) to create the 2-Wasserstein barycenter with Gaussian mapping (*GaussW₂* barycenter hereafter).

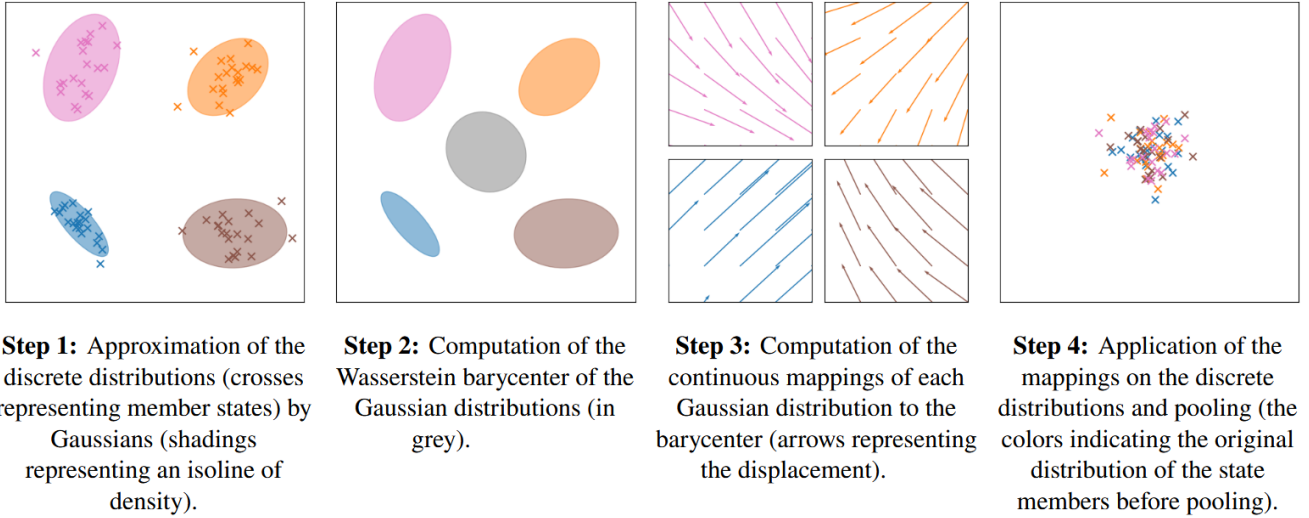


Figure 1. Illustration of the Wasserstein barycenter with Gaussian mapping approach.

This approach computes mappings and barycenters on continuous (multivariate normal) distributions, thereby avoiding variance shrinkage of the barycenters due to limited discrete sample sizes (see Appendix A). Computing the mean and covariance of a discrete distribution for the multivariate normal approximation is less sensitive to the sampling of the distributions (Gnasounou et al., 2023), and regularization can be used if needed to ensure the invertibility of the covariance matrices $\hat{\Sigma}_k$. This is especially important when the number of members (i.e. samples) is small compared to the number of dimensions. The approximation of the discrete distributions by multivariate normal distributions is a strong hypothesis considering that ensembles in climate science are propagated according to nonlinear evolution equations. However, this is a classical hypothesis for modeling stochastic processes and in signal processing. Moreover, this hypothesis is only used to derive Gaussian barycenter and the mappings which are then applied to the original discrete distributions which can be seen as a second-order calibration of the

models onto a barycenter which preserves higher-order moments through the individual samples. For variables with a strongly non-Gaussian distribution (such as precipitation), it is also possible to first apply a transformation to make the distribution more Gaussian (e.g., using Gaussian anamorphosis; see Bertino et al. (2003); Lussana et al. (2021)), compute the barycenter in the transformed space, and then apply the inverse transform.

Another advantage of this approach is the computational cost (in low dimensions). Estimating the parameters of the multivariate normal distribution from the discrete one has a computational complexity of $\mathcal{O}(Nd^2)$ (where N is the number of points in the discrete distribution, and d the number of dimensions, here $d = n_t$ lead times). The Gaussian mappings and Gaussian barycenter computation have a complexity of $\mathcal{O}(d^3)$ and $\mathcal{O}(d^3 \log(1/\epsilon))$ respectively (where ϵ is the tolerance in the fixed-point algorithm). On the other hand, computing the exact barycenter requires multiple resolutions (per model and iteration) of exact optimal transport problems $\mathcal{O}(N^3 \log(N))$ (Agueh and Carlier, 2011).

2.2.3 Illustration of the application of barycenters to ensemble forecasts

The L_2 and W_2 distances have different properties in the distribution space, leading to distinct multi-model ensembles. The L_2 distance focuses on the “vertical” differences between distributions. This means that the L_2 distance between two distributions with disjoint supports is equal to the sum of their L_2 norms, no matter how distant their supports are (e.g., it is equal to $2 \sum_i^N 1/N^2$ for two discrete distributions with each N equiprobable points). As a result, the members of the L_2 barycenter always coincide with the members of the initial ensembles, regardless of the distance between the supports of the input distributions. This means that the members of the L_2 barycenter are physically consistent, according to the numerical weather model they were drawn from. In contrast, the $GaussW_2$ distance is a “horizontal” measure of the difference between two distributions, in the sense that it is based on the horizontal displacement (or mapping) between them (Santambrogio, 2015). This enables the associated $GaussW_2$ barycenter to have a different support than the input distributions. In practice, that means that the members of the $GaussW_2$ barycenter do not coincide with any of the input ensemble members. Although the covariance structures of the inputs are still taken into account, they are merged (step 2 of the method in Section 2.2.2), allowing the barycenter to contain forecasts that could not have been generated by the individual models. The $GaussW_2$ barycenter is thus more flexible, but does not guarantee the same level of physical consistency as the L_2 barycenter.

Figure 2 gives an illustration of the two barycenters applied to the combination of two synthetic ensemble forecasts with $N_1 = 10$ and $N_2 = 15$ members respectively. The members of the L_2 barycenter in Figure 2b correspond to the members of the input ensembles in Figure 2a. The weight of each member of the L_2 barycenter is given by the weight of the corresponding member in the input ensemble ($1/N_1$ or $1/N_2$) multiplied by the corresponding model’s weight (here $\lambda_1 = \lambda_2 = 0.5$). The $GaussW_2$ barycenter shown in Fig. 2c has a different structure compared to the L_2 barycenter. As expected, a major difference is that none of its members belong to the input ensembles. In other words, while the support of the L_2 barycenter is the union of the input supports, the support of the $GaussW_2$ barycenter lies along the path of the optimal transport and can therefore be different. This also allows for the $GaussW_2$ barycenter to retain certain characteristics of the input distributions. For example, in this illustrative case, both input distributions are unimodal, and so is the $GaussW_2$ barycenter. This is not the case for the L_2 barycenter which has two modes, each associated with the mode of one of the input distributions. The two

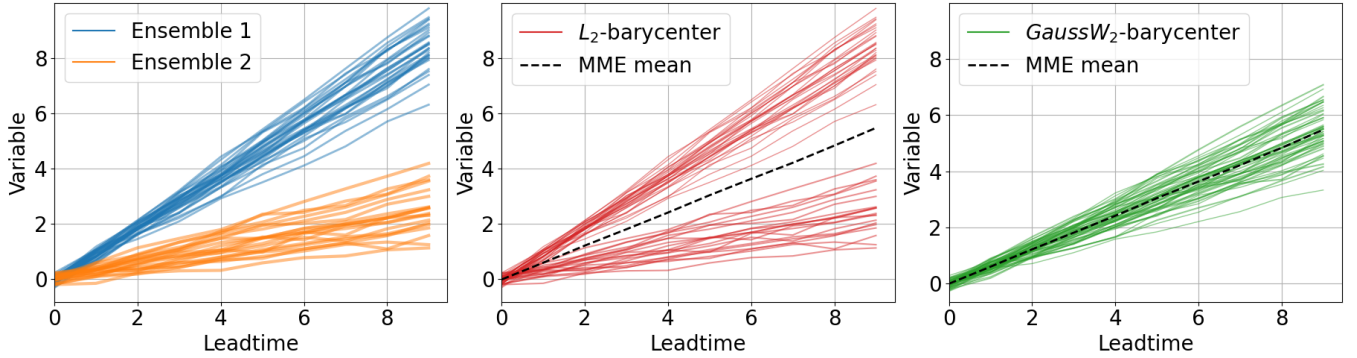


Figure 2. Illustration of (equal-weights) multi-model ensembles using barycenters for a synthetic variable. The ensemble can be seen as a discrete distribution whose points are the time series of each member. The weight of the points in the distribution is indicated here by the thickness of the line.

barycenters may be attractive for different types of forecasts. The $GaussW_2$ barycenter preserves some geometric properties of the input distributions, while the L_2 barycenter retains bimodality. Actual forecasts are more complex, and assessing the different barycenters will require extensive testing with different criteria. The purpose of Figure 2 is to emphasize the (hitherto
230 largely untapped) variety of ways to build multi-model ensemble forecasts.

An important fact to note is that, despite their differences, the two barycenters have the same ensemble means (for the same model weights). Their mean is equal to the weighted mean of the input distributions’s means (this follows easily from Eq. (3) for the L_2 barycenter, and see Peyré and Cuturi (2020, Remark 9.1) for the W_2 and $GaussW_2$ barycenters). Thus, the difference between the barycenters is how they represent the forecast uncertainty. This raises the question: how different are
235 the L_2 and $GaussW_2$ distributions, and which one captures forecast uncertainty better? We address this question empirically for a particular case study in the next sections.

3 Data and methodology

In this study, we focus on subseasonal forecasts over Europe during the boreal winter. We consider one large-scale and two surface variables: the geopotential height at 500 hPa (z500), the daily temperature at 2 m (t2m), and the wind speed at 10 m
240 (w10m). The latter two variables are interesting to predict for the energy sector, as energy demand is particularly dependent on the temperature in winter (due to the use of electrical heating), while the wind speed is an essential variable for estimating wind farm production. The geopotential height at 500 hPa is included as a traditional large-scale variable for comparison.

3.1 Data

To make predictions for this case study and to validate them, we need both a dataset of ensemble forecasts from multiple dynamical models, to which the barycenters will be applied and a reference dataset against which the forecast skill will be evaluated.

3.1.1 S2S data

For this first implementation, we selected four models from the S2S database (Vitart et al., 2017): the European Centre for Medium-Range Weather Forecasts (ECMWF) model, the National Centers for Environmental Prediction (NCEP) model, the Environment and Climate Change Canada (ECCC) model, and the Korea Meteorological Administration (KMA) model. An important criterion in model selection is their diversity. As previously explained, one advantage of using multi-model ensemble is to better sample model error. If the models have similar construction, this error is not sampled properly. Other criteria include time coverage and consistency in production over the study period.

Table 1. Description of the four models from the S2S database (characteristics for the period used). See Vitart et al. (2017) for more details.

	ECMWF	NCEP	ECCC	KMA
Forecasts				
Time range	d 0–46	d 0–44	d 0–32	d 0–60
Resolution	Tco639/Tco319 L137	T126 L64	Yin-Yang grid at 35° (varying vertical resolution)	N216 L85
Frequency	Twice a week (Monday, Thursday)	Daily	Weekly (Thursday)	Daily
Ensemble size (original)	50+1	15+1	20+1	4
Ensemble size after pre-processing	51	48	21	32
Reforecasts				
Method	On the fly	Fixed	On the fly	On the fly
Period	Past 20 years	1999–2010	fixed period (varying)	fixed period (varying)
Frequency	Twice a week	Daily	Weekly	4/month
Ensemble size	10+1	3+1	3+1	3

The forecasts and reforecasts from the selected models were retrieved from the S2S database through the ECMWF’s Meteorological Archival and Retrieval System (MARS) directly on a $1.5^{\circ} \times 1.5^{\circ}$ grid for our study domain shown in Figure 3,

that is Europe ($34^{\circ}N - 74^{\circ}N$, $13^{\circ}W - 40^{\circ}E$). We retrieved the daily mean 2 m temperature, the instantaneous values of the 10 m wind speed, and the 500 hPa geopotential height at midnight (00Z). We selected the forecasts initialized during the winter months (December-January-February, DJF) for the 2016–2022 period, for which the four models have matching initialization dates. This resulted in a total of 90 simulations spanning seven winters. After pre-processing (described below), the forecasts were resampled to weekly means. The barycenter-based MME methods were applied to the weekly time series (i.e., $n_t = 4$ weeks), which was also the temporal resolution used for forecast evaluation.

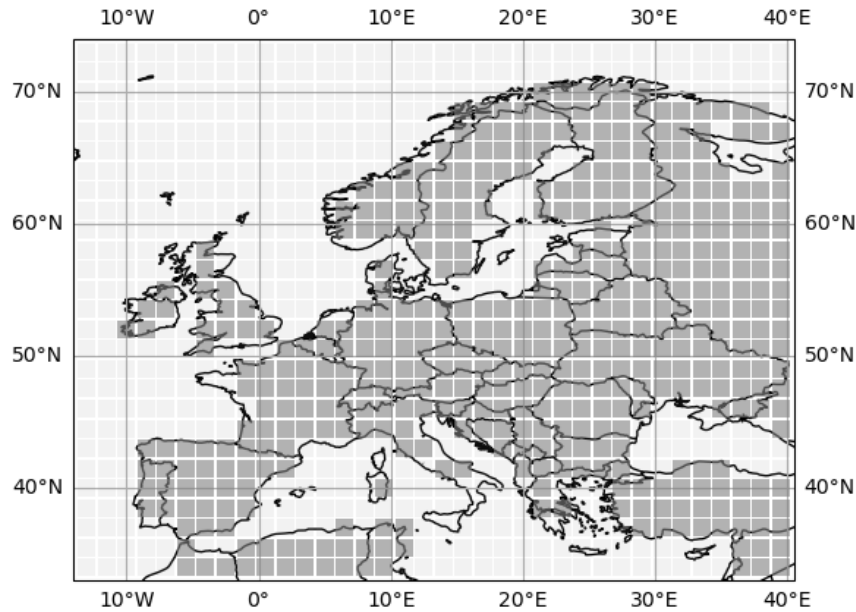


Figure 3. Study domain with the $1.5^{\circ} \times 1.5^{\circ}$ lat/lon grid from the forecasts. Only the land points (in grey) are used in this case study.

Pre-processing:

- The forecasts from the four models are calibrated using the Mean and Variance Adjustment (MVA, Leung et al. (1999); Manzanas et al. (2019)) method, following Goutham et al. (2022). Ensemble forecasts suffer from systematic errors (e.g. mean bias) due to uncertainties in ensemble initialization and model formulation. In particular, forecasts tend to drift away from the observed climate toward the model climatology as lead time increases (Takaya, 2019), making statistical correction essential at extended-range time scales.

The model climatology is estimated from reforecasts, which are “retrospective” forecasts generated with the same model but for past dates. The observed climatology is derived from a reference dataset (see Section 3.1.2 below). The calibration method (MVA here) is then applied to statistically adjust the forecasts. To ensure a robust estimation of the climatologies,

we aggregate all forecasts initialized within a 15-day window, comprising the same calendar day as the forecast starting date and the previous 14 days¹ (Manrique-Suñén et al., 2020). The observed climatology is computed on the same dates for consistency.

- The NCEP and KMA models have smaller ensemble sizes but are produced daily. We thus create a lagged ensemble by aggregating the forecasts with those initialized the two preceding days for NCEP and the seven preceding days for KMA. This allows us to improve their skill and to increase their ensemble size, making them more comparable to the other models (48 and 32 members for NCEP and KMA, respectively after lagging). The ECCC model has a smaller ensemble size than ECMWF and the lagged NCEP and KMA models. However, lagging is not applied to it because of its production frequency (it is only produced weekly).

280 3.1.2 Reference data

The Modern-Era Retrospective analysis for Research and Applications, version 2 (MERRA-2, Gelaro et al. (2017)) reanalysis is used as reference for calibration and forecast validation. We use a reanalysis as reference in order to have a spatially and temporally complete dataset. MERRA-2 was chosen because it is a recent reanalysis covering both the calibration and validation periods and is based on a different circulation model from those used in the selected S2S models. This would not be the case, for example, with the ERA5 reanalysis which is also produced by the ECMWF with a similar model as the S2S forecasts (Hersbach et al., 2020). The three variables of interest were retrieved from NASA Goddard Earth Sciences (GES) Data and Information Services Center (DISC) on MERRA-2's native grid (0.5° lat $\times$ 0.625° lon grid). The data were re-gridded to match the $1.5^\circ \times 1.5^\circ$ lat/lon grid of the forecasts using bilinear interpolation with the Climate Data Operators (CDO, Schulzweida (2022)).

290 We also use MERRA-2 to build a 30-year rolling climatology. The climatology is a common benchmark for forecast validation, and it is also used to compute skill scores (see Section 3.2). For a given day of the year, the climatology consists of MERRA-2 values for the same day over the previous 30 years. The climatology is thus an ensemble with 30 members, each corresponding to a different year.

3.2 Evaluation

295 We now introduce the evaluation scores used in this study. We use the Continuous Ranked Probability Score and related skill scores to evaluate the accuracy of the forecasts, and the Spread-Skill Ratio to investigate their calibration. Due to the low number of samples in our case study (90 simulations), we do not investigate the prediction of extreme events. All scores used in this study are summarized in Table 2.

¹previous 15 days for KMA to better match the frequency of its reforecasts

3.2.1 Continuous Ranked Probability Score (CRPS)

300 The CRPS is a widely used score for probabilistic forecasts of continuous variables (Matheson and Winkler, 1976; Gneiting and Raftery, 2007; Wilks, 2019). The CRPS is actually the squared L_2 distance between the Cumulative Distribution Function (CDF) of the ensemble forecast, and the CDF of the observation:

$$\text{CRPS}_n = \int_{-\infty}^{+\infty} [F_{\text{fc},n}(y) - F_{\text{ref},n}(y)]^2 dy \quad (5)$$

where F_{fc} is the CDF of the forecast, F_{ref} is the CDF of the observation and $n \in \llbracket 1, N \rrbracket$ is the ensemble member index. The
 305 CDFs are computed empirically from the ensembles. In general the reference is deterministic, and its CDF is then a step function $F_{\text{ref}}(y) = \mathbf{1}_{y \geq x_{\text{ref}}}$, where x_{ref} is the deterministic value of the reference.

The accuracy of the ensembles is evaluated by computing the mean CRPS (over the forecasts indexed by n). For easier comparison across variables, we use the Continuous Ranked Probability Skill Score (CRPSS) which compares the mean CRPS of the ensemble forecast to the mean CRPS of the climatology, but statistical tests will be performed on the CRPS distribution.

310 We consider two other scores based on the CRPS. To evaluate robustness to outliers, we use the proportion of skillful forecasts (hereafter CRPSp), which corresponds to the percentage of simulations for which the model has a better CRPS than the climatology (Goutham et al., 2022). We consider that the model has some skill if 50% of the forecasts are skillful. As a counterpart, we introduce a third score that focuses on the tail of the distribution (i.e. forecasts far from the truth). The proportion of critical failures (CRPSf) is defined as the percentage of simulations for which the model has a CRPS_n larger
 315 than twice that of the climatology (this metric is thus negatively oriented, contrary to the CRPSS and CRPSp). The CRPSp and CRPSf should be interpreted together.

Remarks on the links between the CRPS and the L_2 barycenter:

- It is interesting to note that for univariate distributions, the L_2 barycenter is also the barycenter with respect to the energy distance (see appendix B). Thus, since the CRPS is identical to the energy distance in 1D (see Wilks, 2019), we can
 320 expect good CRPS performance for the L_2 barycenter.

For symmetry, the Wasserstein distance could also be used as a score for forecast validation. However, it is equivalent to the RMSE (averaged over the members) in the case of a deterministic observation and so it does not evaluate well the representation of the uncertainties in the forecasts. Thus, we do not use it here.

- The CRPS of the L_2 barycenter $\mu_{L_2}^\alpha$ of the distributions $(\mu_1, \dots, \mu_d)$ (with weights $\lambda_1, \dots, \lambda_d$) can be expressed as a
 325 function of their CRPS with respect to the observation and their cross-CRPS:

$$\text{CRPS}(\mu_{L_2}, \nu_{\text{obs}}) = \sum_{i=1}^n \lambda_i \text{CRPS}(\mu_i, \nu_{\text{obs}}) - \sum_{i=1}^n \sum_{j \neq i} \frac{1}{2} \lambda_i \lambda_j \text{CRPS}(\mu_i, \mu_j) \quad (6)$$

where ν_{obs} is the observation (see appendix C). This expression clearly shows that the CRPS of the L_2 barycenter is always smaller than the average CRPS of the input distributions. In fact, the more different the input distributions are (i.e., large cross-CRPS), the more advantageous it is to merge them.

330 3.2.2 Spread-skill ratio

The CRPS evaluates the accuracy of the forecasts, that is, their overall quality (Wilks, 2019). In order to focus on the calibration (or reliability) of the ensembles, we also look at the spread-skill ratio (SSR) defined as the ratio of the ensemble standard deviation to the root mean square error RMSE of the ensemble mean. A perfectly reliable ensemble has a SSR equal to one, while larger SSR values indicates overdispersion and smaller values underdispersion (Fortin et al., 2014). The SSR can be used
335 as a first evaluation of an ensemble’s reliability (a value of 1 is necessary, though not sufficient, for a fully reliable ensemble).

Table 2. Scores used for evaluation.

Name	Formula	Range	Comment
$\overline{\text{CRPSm}_l}$	$\frac{1}{N_{for}} \sum_{k=1}^{N_{for}} \langle \text{CRPS}_{fc,k,l} \rangle$	$[0, +\infty)$	negatively oriented
CRPSS	$1 - \frac{\overline{\text{CRPSm}_l}(fc)}{\overline{\text{CRPSm}_l}(clim)}$	$(-\infty, 1]$	positively oriented
CRPSp	$\frac{1}{N_{for}} \sum_{k=1}^{N_{for}} \frac{100}{N_{lat} N_{lon}} \sum_{i=1}^{N_{lat}} \sum_{j=1}^{N_{lon}} [\text{CRPS}_{fc,k,i,j,l} < \text{CRPS}_{clim,k,i,j,l}]$	$[0, 100]$	positively oriented
CRPSf	$\frac{1}{N_{for}} \sum_{k=1}^{N_{for}} \frac{1}{N_{lat} N_{lon}} \sum_{i=1}^{N_{lat}} \sum_{j=1}^{N_{lon}} [\text{CRPS}_{fc,k,i,j,l} > 2 \text{CRPS}_{clim,k,i,j,l}]$	$[0, 100]$	negatively oriented
SSR	$SSR = \frac{\sqrt{\frac{1}{N_{for}} \sum_{k=1}^N \langle \text{var}_m(\text{fc}_{k,m}) \rangle}}{\sqrt{\frac{1}{N_{for}} \sum_{k=1}^N \langle (\text{mean}_m(\text{fc}_{k,m}) - \text{obs}_k)^2 \rangle}}$	$[0, +\infty)$	SSR > 1: overdispersion SSR < 1: underdispersion
Notations	$i \in \llbracket 1, N_{lat} \rrbracket$: latitude; $j \in \llbracket 1, N_{lon} \rrbracket$: longitude; $l \in \llbracket 1, N_{day} \rrbracket$: lead time; $k \in \llbracket 1, N_{for} \rrbracket$: forecast (i.e. starting date); $m \in \llbracket 1, N_{mem} \rrbracket$: member.		
Remarks	- $\langle \cdot \rangle$ denote the area-weighted spatial average. - mean_m and var_m are the mean and the variance over the ensemble members respectively.		

4 Results

In this section, we evaluate the combination of four models’ ensembles applying the two barycenter-based MME methods described in Section 2 to the case study described in Section 3. All the results shown here are obtained for barycenters with equal weights (that is $\lambda_{ECMWF}, \lambda_{NCEP}, \lambda_{ECCC}, \lambda_{KMA} = 1/4$). We tested optimizing the weights with respect to the CRPS, but it

340 leads to either little improvement or slight deterioration of the forecast skill depending on the variables and scores (not shown here).

4.1 Validation and comparison of the barycenter-based MME

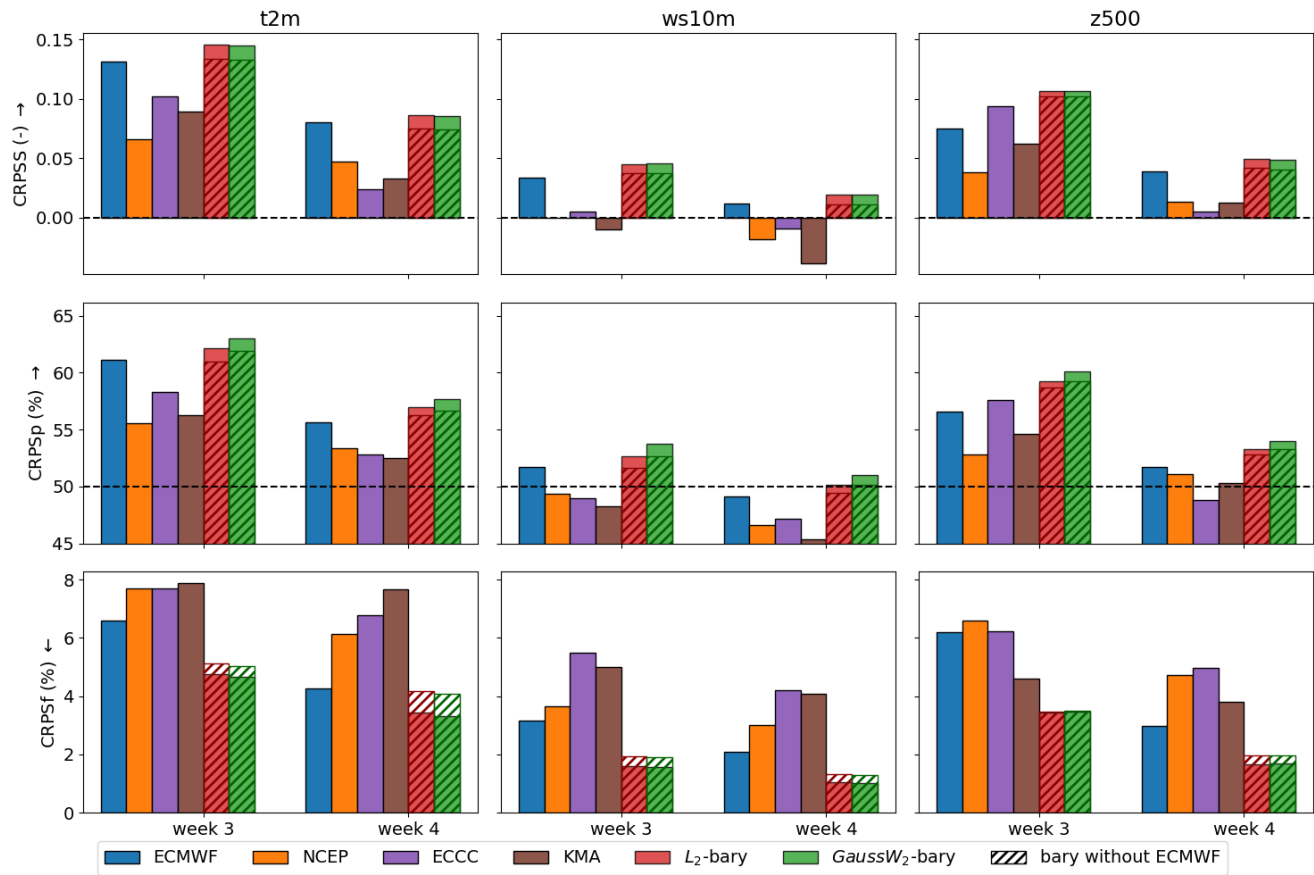


Figure 4. Skill scores for the weekly 2m temperature (left), 10m wind speed (center) and geopotential height at 500hPa (right): the CRPSS (top), the proportion of skillful forecasts (CRPSp, middle) and proportion of critical failure (CRPSf, bottom).

Figure 4 summarizes the performance of the four SMEs and the two barycenter-based MMEs. For the latter, the bars represent the performance of the barycenters combining the four SMEs (the hatched areas represent the performance of the barycenters of three SMEs out of four; these results are discussed below, Sect. 4.3). The distribution of the scores (CRPSS, CRPSp and CRPSf) over the starting dates is represented by its mean in the plot, and is used to test if the models' scores are significantly different from each other at a 5% significance level. The significance is estimated using the Wilcoxon signed-rank test, a non-parametric paired statistical test (Wilcoxon, 1945). We chose to show the CRPSS here instead of the mean CRPS for an

easier comparison across variables. However, the mean and the statistical tests are computed on the (spatially averaged) CRPS distributions. The p-values of these statistical tests are shown in Appendix D.

For all three variables and three scores, the two barycenters outperformed the SMEs, however the gain with respect to the best performing SME is sometimes small. In particular, ECMWF has the best performance among the four SMEs for the surface variables, the 2 m temperature and the 10 m wind speed. In terms of CRPSS, the two barycenters have similar values and are significantly different from all the SMEs for the 10 m wind speed and the geopotential height (see Tables D2 and D3). Even though two or three of the models (depending on the week) have a negative CRPSS for the 10 m wind speed, merging them with more skillful SMEs still adds value: the barycenters have a significantly better CRPSS. For the 2 m temperature, the barycenters do not have a significantly different CRPSS than ECMWF, which already has a relatively large CRPSS (see Table D1).

In terms of CRPSp, the *GaussW₂* barycenter is significantly better than the others, including the *L₂* barycenter that ranks second (Tables D1, D2 and D3). That is, the barycenters, and especially the *GaussW₂* barycenter, are more often better than the climatology compared to the SMEs. The barycenter-based ensembles also significantly reduce the risk of large failures as shown by their lower CRPSf. The two barycenters have similar CRPSf values, clearly lower than the values of the SMEs. In particular, the proportion of critical failures is divided by almost two for the 10 m wind speed compared to the best SME, ECMWF.

Variations in scores across the spatial domain for each model are large, in the sense that they tend to dominate over differences between models (not shown here). We thus expect this to be also the case for the MMEs. Figure 5 shows the differences between ECMWF and the *GaussW₂* barycenter's scores. We chose ECMWF as a reference, since it has the best performance among the SMEs for two of the three variables (the surface ones). The map of the differences between the *L₂* barycenter and ECMWF is very similar, and so is not shown here. For the CRPSS, one can see clear areas that benefit from the MMEs, but these areas are not the same depending on the variables. For the 2 m temperature, the barycenter is more advantageous in the North-West, above the North Sea, Atlantic Ocean and Scandinavia. ECMWF is better above Sweden and Finland for the 10 m wind speed. For this variable, the barycenter benefits more in central Europe as well as southern and eastern Europe. The spatial pattern is smoother for the geopotential height, with the largest differences above northern Europe, where the barycenter is more skillful. Despite these spatial variations across the three variables, the barycenter outperforms ECMWF at a majority of grid points.

The maps for the CRPSp differences are more pixelated, however the positive and negative areas roughly match those of the CRPSS differences. In contrast, the maps of the CRPSf differences show very different spatial patterns. For example, ECMWF performs better over Spain for the two surface variables for both the CRPSS and the CRPSp, but the barycenter has a better CRPSf there. In fact, the barycenter outperforms ECMWF in terms of CRPSf almost everywhere. This means that combining ensemble forecasts has a stabilizing effect, avoiding extremely bad forecasts. However, the differences are relatively small, less than 3% at most grid points.

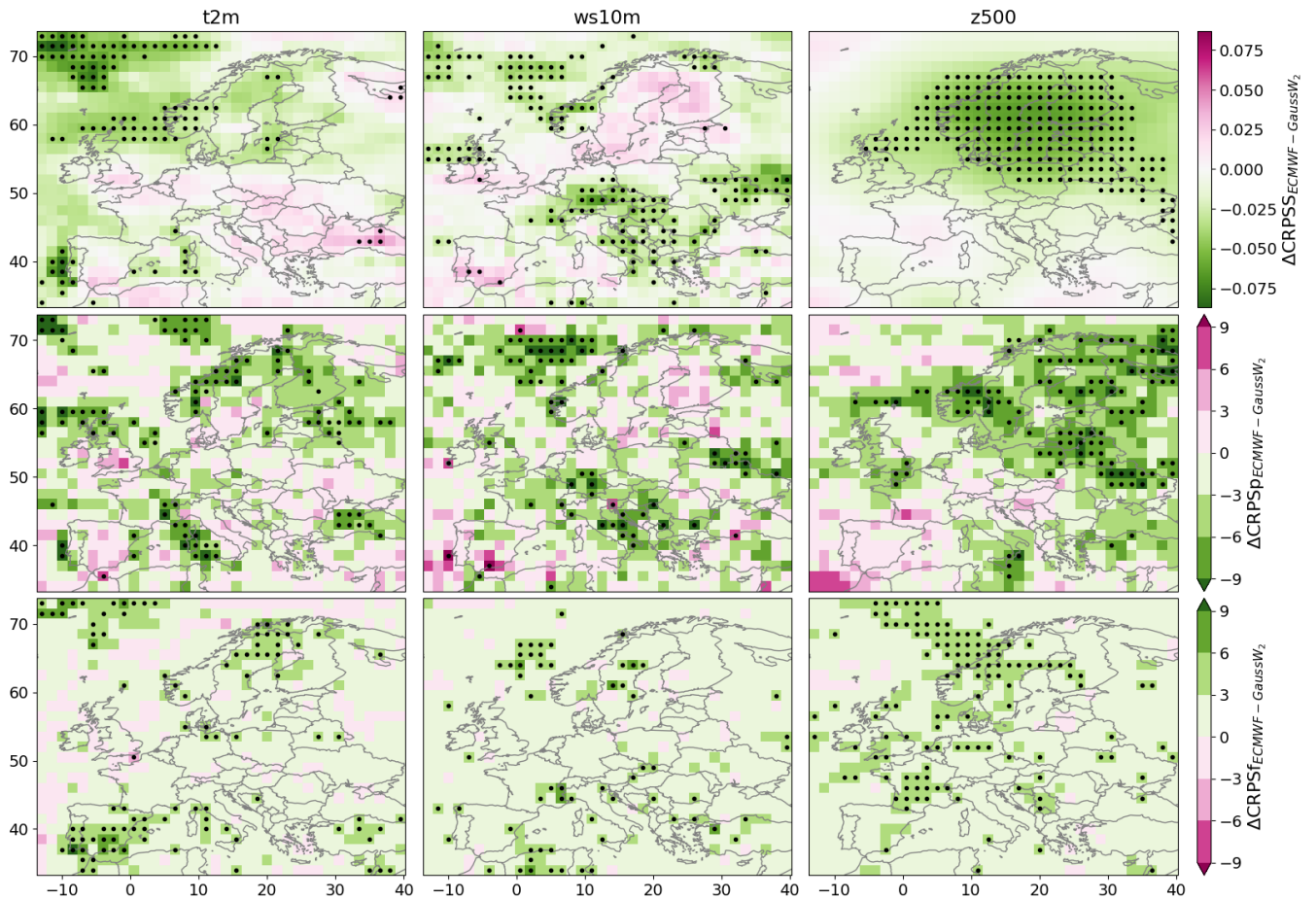


Figure 5. Maps of the differences between ECMWF and $GaussW_2$ barycenter in terms of (top) CRPSS, (middle) CRPSp, and (bottom) CRPSf averaged over weeks 3 and 4. Green indicates that $GaussW_2$ barycenter performs better and pink that ECMWF does. For readability reason, a discrete colorbar is used for the $\Delta CRPSp$ and $\Delta CRPSf$ which are more pixelized. Dots indicate grid points where the difference is statistically significant at the 5% level (using a two-sided Wilcoxon signed rank test for the CRPS and a McNemar test for the CRPSp and CRPSf).

4.2 Ensembles calibration

In order to get insight into the calibration of the ensemble forecasts, we look at their SSR maps in Figure 6. The spatial patterns of weeks 3 and 4 being similar, only their average is shown here. The spatial patterns are consistent across the SMEs. They tend to be under- or over-confident in the same regions, but with different intensities. For example, all models are over-dispersed over Scandinavia for the 2m temperature forecasts, with KMA having the least and ECCC the most overdispersion. On the contrary, they are all under-dispersed over the Mediterranean Sea.

Note that, by construction, the barycenters have the same ensemble mean and so the same RMSE, so that the differences in their SSR are only due to their spread. Their SSRs have similar patterns as the SMEs. However, one can notice that the L_2 barycenter tends to have more spread than the *GaussW*₂ barycenter: it shows more overdispersion and less underdispersion. This can be explained by the way they were constructed. The variance of the L_2 barycenter is indeed given by

$$\text{Var}(\mu_{L_2}) = \sum_{i=1}^d \lambda_i \text{Var}(\mu_i) + \sum_{i=1}^k \lambda_i \left(\bar{\mu}_i - \sum_{j=1}^k \lambda_j \bar{\mu}_j \right)^2, \quad (7)$$

where $\bar{\mu}_i$ is the mean of the distribution μ_i . The variance of the barycenter is thus always equal to or higher than the average of the SME's variances. This property is advantageous in the case of under-dispersed input ensembles. However, it is a disadvantage if they are better calibrated or already over-dispersed (e.g. Zhou et al., 2022; Eade et al., 2014). In those cases, the *GaussW*₂ barycenter is more advantageous. In order to build it, the input ensembles are first centered and their covariances scaled before being pooled together. Thus, its covariance is the one of the Gaussian barycenter (Sect. 2.2.2).

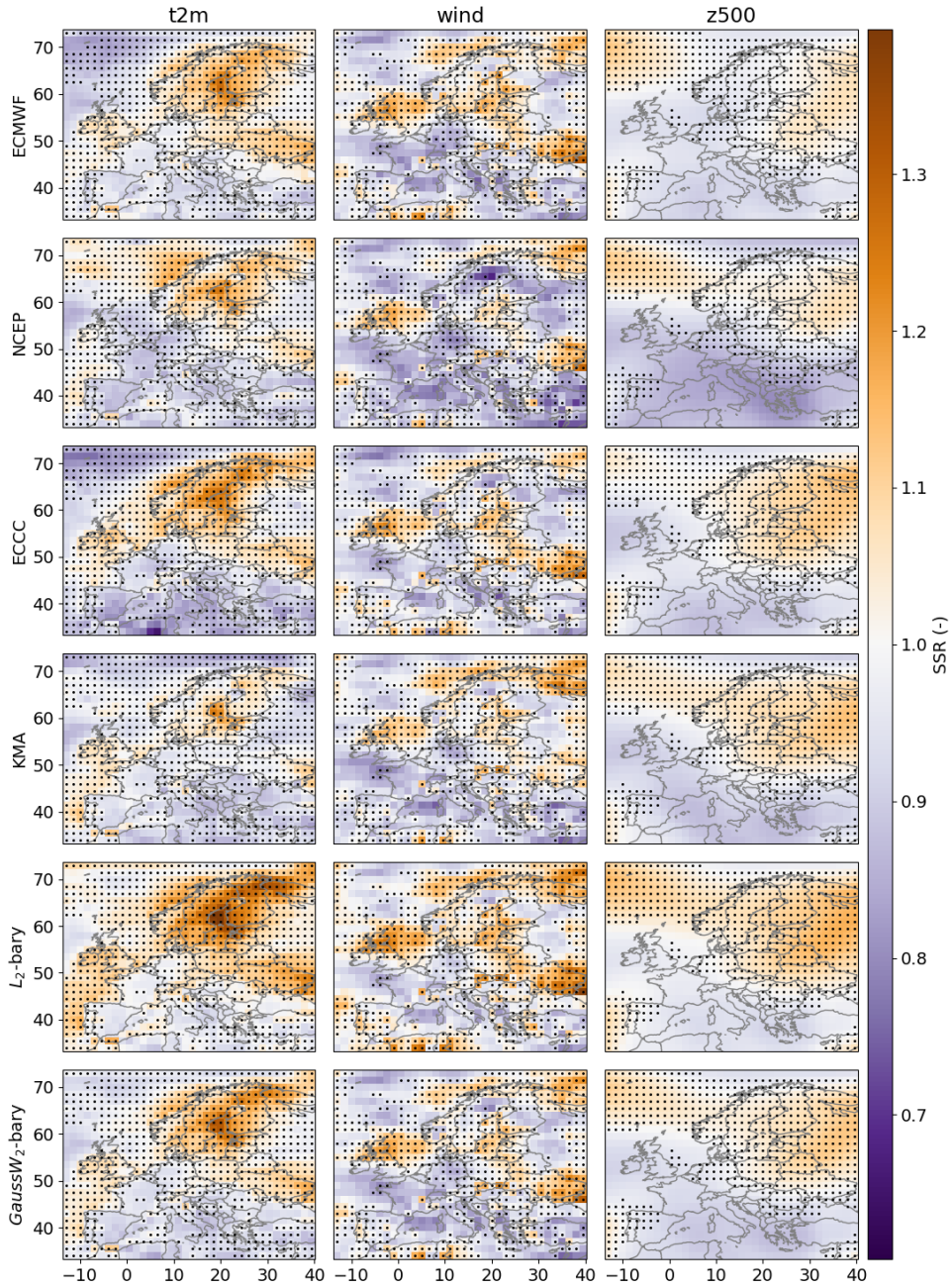


Figure 6. Spread-Skill-Ratio (SSR) for the weekly 2m temperature (left), 10m wind speed (center) and geopotential height at 500hPa (right) averaged over weeks 3 and 4 for the four single-model ensembles and the two barycenters. Orange indicates overdispersion (i.e. under-confident forecasts), and blue under-dispersion (i.e. overconfident forecasts). Dots indicate grid points where the ensemble variance and MSE of the ensemble mean are statistically different at the 5% level (using a two-sided Wilcoxon signed rank test on the distribution of variance and MSE).

4.3 MME without ECMWF

In Section 4.1, ECMWF is found to be more skillful than the other three SMEs for the two surface variables. Moreover, although the two barycenters have overall better performance, ECMWF is still better in some regions. Thus, to investigate the importance of ECMWF within the combination, we compute the barycenters of NCEP, ECCC and KMA (without ECMWF). They are compared to the barycenters with ECMWF in Figure 4 (hatched bars). As expected, we observe a decrease in performance for all scores and variables. This is also true for most grid points when looking at the maps of differences except for those points where the barycenters (with four models) are notably more skillful than ECMWF (not shown here).

When comparing the barycenters with the three SMEs used to build them, we can make similar observations as in Section 4.1. Specifically, the L_2 and $GaussW_2$ barycenters have similar CRPSS and CRPSf, but the $GaussW_2$ barycenter has a significantly better CRPSp (at the 5% level, Tables D4, D5 and D6). Both barycenters also outperform the three SMEs for all metrics. However, two features are noteworthy. First, the CRPSS and CRPSp differences between the barycenters and the SMEs for the surface variables are notably larger than those observed between the barycenters with four models and ECMWF. A particularly striking example is the CRPSS of the 10 m wind speed: while the three SMEs show little to no skill, with very low or even negative CRPSS, their barycenters have CRPSS similar to ECMWF's. Second, the barycenters with only three models are performing as well as ECMWF in terms of CRPSS and CRPSp (there are no significant statistical differences, see Appendix D2) and better in terms of CRPSf. Thus, with three less skillful models, one can reproduce the considerably better performance of ECMWF in the case studied here.

5 Discussion

5.1 L_2 barycenter versus $GaussW_2$ barycenter

We have seen that the MMEs improve the average forecast skill compared to the SMEs with respect to all scores and for all the variables considered, even though a SME may perform better at some locations. In general, the two barycenters have similar CRPSS and CRPSf, but the $GaussW_2$ barycenter has a better CRPSp. In other words, the two barycenters are similar on average, but the $GaussW_2$ is better more often without leading to an increase in very poor forecasts (with respect to the CRPS). Another important difference between the barycenters is the way they represent forecast uncertainty through their ensemble spreads. The L_2 barycenter always has a larger ensemble spread than the $GaussW_2$ barycenter, which in turn has an impact on their reliability.

The two barycenters have the same means (which only depend on the means of the input distributions), but their covariances differ. The covariances of the Wasserstein barycenter depend solely on the covariances of the input ensembles (see Section 2), while the covariances of the L_2 barycenter depend on both their means and their variances. If we assume that the variances of the input ensembles represent the uncertainty due to their initial conditions, their average in the formula of the L_2 barycenter variance (first term of Eq. (7)) can be interpreted as a measure of the MME's initial-condition uncertainty. The second term, the variance of the input ensemble means, can be seen as a measure of model uncertainty as it quantifies the differences

430 between ensemble (means). Thus, the variance of the L_2 barycenter accounts for model errors but ignores differences in higher moments than the ensemble mean. On the other hand, the *GaussW*₂ barycenter merges the forecast uncertainties of the different models (i.e. the uncertainty due to the initial conditions). Instead, information about model uncertainty lies in the value of the barycenter distance: $\min_{\mu} \sum_{k=1}^d \lambda_k \cdot d(\mu_k, \mu)^2$. This information is not exploited here, but it could be considered to account for model uncertainty. Finally, not only are dynamical and model uncertainties captured by different quantities
 435 for the two barycenters (variance of the barycenter versus barycenter distance), but the way they are measured also differs. Understanding the implications of these differences requires further investigation.

5.2 Best model versus models combination

There can be several reasons why merging ECMWF with less skillful models leads to barycenters with improved performance. First, the three SMEs have generally less skill than ECMWF, but can be occasionally better for some given locations, lead
 440 times, or initialization dates. The barycenters can exploit this information to improve skill thanks to error cancellation and to the non-linearity of the skill metrics (Hagedorn et al., 2005). Second, the ECMWF forecasts may have good performance but be overconfident. In that case, adding other models with lower skill increases the spread of the ensemble and can move the ensemble mean towards the truth, as shown by Weigel et al. (2008) for seasonal ensemble forecasts. For the L_2 barycenter, this can be seen from Equation (6) which shows that the CRPS of the L_2 barycenter is composed of two parts: the weighted
 445 average of the CRPS of the SMEs minus the weighted average of the CRPS between all pairs of SMEs. Thus, if one model has a worse CRPS than another, it can still improve the CRPS of the barycenter if the CRPS between the models is large enough to compensate for its own CRPS.

5.3 Weighted barycenters

Equal weighting of the models in the pooling method is the simplest and most used approach. However, some studies investi-
 450 gate the use of weighted multi-model ensembles with mixed results (at the weather or seasonal scale in Weigel et al. (2008); Casanova and Ahrens (2009); Kharin and Zwiers (2002) and at the climate scale in Haughton et al. (2015)). At the subseasonal scale, Wanders and Wood (2016) used multivariable linear regression on the ensemble means to derive the multi-model weights. They show that the weighted multi-model ensemble has better deterministic performance but also better probabilistic performance (in terms of the Brier score) than the unweighted multi-model ensemble.

455 Here, we have shown results using equal weights for the models, but the barycenter formulation easily allows for the construction of weighted MMEs. Results for weights that do not depend on space or time, minimizing the mean CRPS, were also obtained but not shown because, contrary to Wanders and Wood (2016), no significant improvement was found. One explanation could be the limited time period used here. The collection of forecasts by the S2S database started in 2015. We thus have less than ten years of data, which may not be enough to find stable weights, as shown by Wanders and Wood (2016);
 460 Kharin and Zwiers (2002). Using reforecasts (instead of forecasts) would allow one to extend the time period. However, it would still be limited to twelve years (due to the limited NCEP reforecast period) and would lead to additional difficulties. In particular, the starting dates of the reforecasts for the different models in the S2S database do not match. In addition, the

reforecasts have fewer members than the forecasts, so the reforecast results may not be transferred to the forecasts without additional assumptions.

465 Yet, another strategy could be to allow the weights to depend on space. We indeed saw that even if the barycenters tend to be more skillful than the SMEs, they can be outperformed by ECMWF for some scores at some locations. This suggests that giving more weight to ECMWF in these regions would be beneficial. Similarly, even if the barycenter without ECMWF showed a general degradation of the scores compared to the barycenters with the four models, it did perform better in some specific regions. Thus, the barycenter with four models would benefit from giving less or no weight to ECMWF in these regions.

470 However, optimizing weights grid point by grid point requires more data and may lead to spatial inconsistencies (Wanders and Wood, 2016; DelSole et al., 2013). An alternative would be to learn weights per region (such as in Wanders and Wood (2016)) or to enforce some spatial smoothness on the weights (also acting as regularization and helping with overfitting).

6 Conclusions

We explore methods to combine ensemble forecasts from multiple models based on barycenters of forecast ensembles. Building

475 on the recognition of the relevance of probabilistic forecasts for S2S prediction, we work directly in the probability distribution space. That is, the ensemble forecasts are manipulated as discrete probability distributions. This allows us to use existing tools from this space and, in particular, the notion of barycenters. Here, we explore two barycenters based on different metrics: the L_2 distance and the Wasserstein ($GaussW_2$) distance. We show that the L_2 barycenter is in fact equivalent to the well-known pooling method and compare it to the new $GaussW_2$ barycenter-based method. Moreover, the variance of the L_2 barycenter

480 can be decomposed into a term representing the average uncertainty of individual forecasts and a term representing model systematic error, while the variance of the $GaussW_2$ barycenter does not account for the latter. Instead, this information is captured by the values of the $GaussW_2$ barycenter.

This first application of our framework to S2S prediction is illustrated through the combination of four single-model ensembles to predict winter surface temperature, surface wind speed, and 500 hPa geopotential height over Europe. By construction,

485 the two barycenters have the same ensemble means, which means they have the same performance in terms of deterministic scores. However, they differ in the way they represent forecast uncertainty. In particular, by construction, the L_2 barycenter is likely to be underconfident due to overdispersion, implying that the L_2 barycenter is advantageous in the case of under-dispersive input ensembles, but disadvantageous in terms of calibration otherwise. More generally, we show that both barycenters perform similarly on average but that the $GaussW_2$ barycenter performs better more often (across the different grid points

490 and dates) with respect to the CRPS. Even in cases where one of the SMEs has superior skill compared to the others (ECMWF for the surface variables), it is still advantageous to combine them into a barycenter. This confirms the interest of multi-model methods shown by previous studies. We also show that by using the other three models, one can build barycenter-based MMEs that are as skillful as the best SME.

This study is a proof of concept to develop the framework and investigate the properties of the barycenter-based MMEs.

495 These results constitute a promising first step towards improving S2S predictions using barycenters to merge ensemble fore-

casts. A next step would be to further investigate the optimization of the weights in the barycenter in order to build weighted MMEs. Another interesting question concerns the possibility of building multivariate Wasserstein barycenters. The L_2 barycenter does not take into account the covariances between the variables, whereas the Wasserstein barycenter does. Finally, the implications of the fact that both barycenters account differently for forecast and model uncertainty deserve further investigation.

500 *Code and data availability.* The Python code for the barycenter-based MMEs is openly available on Github (https://github.com/clecoz/OT_for_MME) and Zenodo (<https://zenodo.org/records/15058503>, Le Coz et al. (2025a)).

This work is based on S2S data. S2S is a joint initiative of the World Weather Research Programme (WWRP) and the World Climate Research Programme (WCRP). The original S2S database is hosted at ECMWF as an extension of the TIGGE database (<https://apps.ecmwf.int/datasets/data/s2s/>). MERRA-2 data is available from the Goddard Earth Sciences Data and Information Services Center (GES
505 DISC) data archive (see Global Modeling and Assimilation Office (GMAO), 2015a, b, c). The pre-processed data are shared on Zenodo (<https://doi.org/10.5281/zenodo.15038871>, Le Coz et al. (2025b)).

Appendix A: Variance shrinkage in discrete Wasserstein barycenter

A potential problem of the discrete Wasserstein barycenter is its reduced variance. Indeed, computing the Wasserstein barycenter directly on discrete probability distributions leads to a variance shrinkage problem which worsens with increasing space
510 dimension and decreasing distribution sampling. This problem can be illustrated with the following empirical experiment:

1. Create two discrete distributions by randomly drawing n samples from two normal distributions with unit-covariance in $\mathbb{R}^d$.
2. Compute the Wasserstein barycenter of the two discrete distributions (using the closed-form expression).
3. Compute the normalized trace of the covariance of the barycenter divided by the number of variables (that is, the average
515 variance of the variables).
4. Repeat the three precedent steps 100 times.

We repeated this experiment for different sampling sizes n and number of dimensions d . The results of this experiment are shown in Figure A1. By construction, the Wasserstein barycenter of the two normal distributions with unit covariance also has a unit-variance (and so an average variance of one). However, one can observe that the discrete barycenter has a lower
520 average variance. The average variance shrinks as the sample size decreases and the number of dimensions increases. For four dimensions and 50 samples, we observe a loss of more than 10% of the average variance. This setting is comparable to our application, where the number of dimensions d corresponds to the number of lead times (i.e. four weeks), and the number of samples n corresponds to the ensemble size (ranging from 21 to 51 members, see Section 3.1.1).

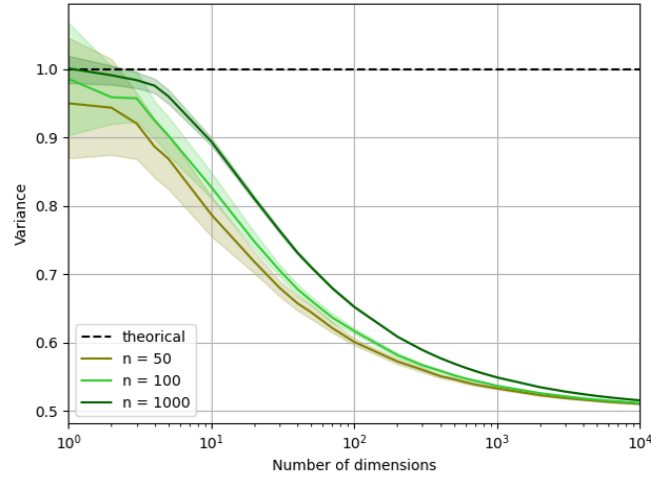


Figure A1. Average variance of the discrete Wasserstein barycenter as a function of the dimension d for different sample size n . The full line represents this variance averaged over the 100 iterations with the minimum and maximum values indicated by the envelop.

Appendix B: Energy distance and its associated barycenter

525 Let μ_1 and μ_2 be two distributions, and F_1 and F_2 be their CDF. That is, $F_1(x) = \int_{-\infty}^x \mu_1(t)dt$ and $F_2(x) = \int_{-\infty}^x \mu_2(t)dt$ $\forall x \in \mathbb{R}$. The squared energy distance between μ_1 and μ_2 is

$$\mathcal{E}^2(\mu_1, \mu_2) = \int_{-\infty}^{+\infty} [F_1(x) - F_2(x)]^2 dx = \int_{-\infty}^{+\infty} \left[\int_{-\infty}^x \mu_1(t)dt - \int_{-\infty}^x \mu_2(t)dt \right]^2 dx$$

The energy barycenter of μ_1 and μ_2 is the solution of the following minimization problem $\mu_{\mathcal{E}}^{\alpha} = \arg \min_{\mu} \alpha \mathcal{E}^2(\mu, \mu_1) + (1 - \alpha) \mathcal{E}^2(\mu, \mu_2)$. Let $\mathcal{B}(\mu) = \alpha \mathcal{E}^2(\mu, \mu_1) + (1 - \alpha) \mathcal{E}^2(\mu, \mu_2)$, then we have:

$$\begin{aligned}
530 \quad \frac{d}{d\mu} \mathcal{B}(\mu) &= \frac{d}{d\mu} \left[\alpha \int_{-\infty}^{+\infty} [F(x) - F_1(x)]^2 dx + (1 - \alpha) \int_{-\infty}^{+\infty} [F(x) - F_2(x)]^2 dx \right] \\
&= \frac{d}{d\mu} \int_{-\infty}^{+\infty} \left(\alpha \left[\int_{-\infty}^x \mu(t) dt - \int_{-\infty}^x \mu_1(t) dt \right]^2 + (1 - \alpha) \left[\int_{-\infty}^x \mu(t) dt - \int_{-\infty}^x \mu_2(t) dt \right]^2 \right) dx \\
&= \int_{-\infty}^{+\infty} \frac{d}{d\mu} \left(\alpha \left[\int_{-\infty}^x (\mu(t) - \mu_1(t)) dt \right]^2 + (1 - \alpha) \left[\int_{-\infty}^x (\mu(t) - \mu_2(t)) dt \right]^2 \right) dx \\
&= \int_{-\infty}^{+\infty} \left(2\alpha \int_{-\infty}^x [\mu(t) - \mu_1(t)] dt \mu(x) \right) dx + \int_{-\infty}^{+\infty} \left(2(1 - \alpha) \int_{-\infty}^x [\mu(t) - \mu_2(t)] dt \mu(x) \right) dx \\
&= 2 \int_{-\infty}^{+\infty} \left(\int_{-\infty}^x (\mu(t) - [\alpha \mu_1(t) + (1 - \alpha) \mu_2(t)]) dt \right) \mu(x) dx
\end{aligned}$$

535 Thus,

$$\frac{d}{d\mu} \mathcal{B}(\mu_{L_2}^{\alpha}) = 0$$

where $\mu_{L_2}^{\alpha} = \alpha \mu_1 + (1 - \alpha) \mu_2$ is the L_2 barycenter. The L_2 barycenter is also a barycenter for the energy distance in 1D.

Appendix C: CRPS and L_2 -barycenter

Let $\mu_1, \dots, \mu_d$ be d probability distributions, and $F_1, \dots, F_d$ their respective cumulative density functions. Their weighted- L_2 barycenter is the probability distribution $\mu_{L_2}^{\lambda}$ given by

$$\mu_{L_2}^{\lambda} = \sum_{i=1}^n \lambda_i \mu_i \tag{C1}$$

where $\lambda = (\lambda_1, \dots, \lambda_n)$ are the barycentric weights such that $\sum_{i=1}^n \lambda_i = 1$.

Let ν_{obs} be the probability distribution of the truth or reference, and F_{obs} its cumulative density function. In the case of a deterministic observation y_r , $\nu_{obs} = \delta_{y_{obs}}$ is a Dirac and F_{obs} a step function.

545 The CRPS of two probability distributions is defined as the squared L_2 -distance between their CDFs:

$$\text{CRPS}(\mu_i, \mu_j) = \int_{-\infty}^{+\infty} [F_i(u) - F_j(u)]^2 du \tag{C2}$$

Thus, we have

$$\begin{aligned}
\text{CRPS}(\mu_{L_2}^\lambda, \nu_{obs}) &= \int_{-\infty}^{+\infty} [F_{L_2}^\lambda(u) - F_{obs}(u)]^2 du \\
&= \int_{-\infty}^{+\infty} \left[\sum_{i=1}^n \lambda_i F_i(u) - F_{obs}(u) \right]^2 du \\
&= \int_{-\infty}^{+\infty} \left[\sum_{i=1}^n \lambda_i^2 F_i(u)^2 + F_{obs}(u)^2 + \sum_{i=1}^n \sum_{j \neq i} \lambda_i \lambda_j F_i(u) F_j(u) - 2 \sum_{i=1}^n \lambda_i F_i(u) F_{obs}(u) \right] du \\
&\quad \text{knowing that } \sum_{i=1}^n \lambda_i = 1 \Rightarrow \lambda_i = 1 - \sum_{j \neq i} \lambda_j \Rightarrow \lambda_i^2 = \lambda_i (1 - \sum_{j \neq i} \lambda_j) = \lambda_i - \lambda_i \sum_{j \neq i} \lambda_j \\
&= \int_{-\infty}^{+\infty} \left[\sum_{i=1}^n \lambda_i F_i(u)^2 + \sum_{i=1}^n \sum_{j \neq i} \lambda_i \lambda_j F_i(u)^2 + \sum_{i=1}^n \lambda_i F_{obs}(u)^2 \right. \\
&\quad \left. + 2 \sum_{i=1}^n \sum_{j \geq i} \lambda_i \lambda_j F_i(u) F_j(u) - 2 \sum_{i=1}^n \lambda_i F_i(u) F_{obs}(u) \right] du \\
&= \int_{-\infty}^{+\infty} \left[\sum_{i=1}^n \lambda_i [F_i(u)^2 + F_{obs}(u)^2 - 2 F_i(u) F_{obs}(u)] \right. \\
&\quad \left. - \sum_{i=1}^n \sum_{j \geq i} \lambda_i \lambda_j [F_i(u)^2 + F_j(u)^2 - 2 F_i(u) F_j(u)] \right] du \\
&= \sum_{i=1}^n \lambda_i \left(\int_{-\infty}^{+\infty} [F_i(u) - F_{obs}(u)]^2 du \right) - \sum_{i=1}^n \sum_{j \geq i} \lambda_i \lambda_j \left(\int_{-\infty}^{+\infty} [F_i(u) - F_j(u)]^2 du \right) \\
&= \sum_{i=1}^n \lambda_i \text{CRPS}(\mu_i, \nu_{obs}) - \sum_{i=1}^n \sum_{j \geq i} \lambda_i \lambda_j \text{CRPS}(\mu_i, \mu_j) \\
\text{CRPS}(\mu_{L_2}^\lambda, \nu_{obs}) &= \sum_{i=1}^n \lambda_i \text{CRPS}(\mu_i, \nu_{obs}) - \sum_{i=1}^n \sum_{j \neq i} \frac{1}{2} \lambda_i \lambda_j \text{CRPS}(\mu_i, \mu_j) \tag{C3}
\end{aligned}$$

Appendix D: Spatial performance's significance

D1 Barycenters with four models

Table D1. Two-sided Wilcoxon test’s p-value for the 2m temperature (t2m) for weeks 3 and 4 combined. Bold indicates that the models are significantly different from each other at the 5% significance level (i.e. p-values < 0.05).

CRPSmean						
	ECMWF	NCEP	ECCC	KMA	L_2 -bary	$GaussW_2$ -bary
L_2 -bary	3.5e-01	3.5e-09	3.3e-09	5.8e-09	-	3.6e-01
$GaussW_2$ -bary	3.4e-01	2.6e-09	2.8e-09	1.1e-08	3.6e-01	-
CRPSp						
	ECMWF	NCEP	ECCC	KMA	L_2 -bary	$GaussW_2$ -bary
L_2 -bary	9.0e-02	4.4e-10	2.0e-09	1.7e-12	-	6.5e-16
$GaussW_2$ -bary	2.5e-03	2.3e-12	5.9e-12	7.1e-15	6.5e-16	-
CRPSf						
	ECMWF	NCEP	ECCC	KMA	L_2 -bary	$GaussW_2$ -bary
L_2 -bary	1.1e-04	6.9e-17	1.8e-20	2.0e-17	-	3.9e-05
$GaussW_2$ -bary	2.1e-05	3.4e-18	8.9e-21	5.9e-18	3.9e-05	-

Table D2. Two-sided Wilcoxon test’s p-value for the 10m wind speed (ws10m) for weeks 3 and 4 combined. Bold indicates that the models are significantly different from each other at the 5% significance level (i.e. p-values< 0.05).

CRPSmean						
	ECMWF	NCEP	ECCC	KMA	L_2 -bary	$GaussW_2$ -bary
L_2 -bary	2.5e-02	1.2e-16	6.8e-12	6.7e-18	-	7.2e-01
$GaussW_2$ -bary	2.0e-02	2.3e-16	1.1e-11	5.5e-18	7.2e-01	-
CRPSp						
	ECMWF	NCEP	ECCC	KMA	L_2 -bary	$GaussW_2$ -bary
L_2 -bary	1.5e-02	6.3e-12	1.7e-11	1.5e-14	-	1.1e-22
$GaussW_2$ -bary	1.3e-05	9.7e-17	3.6e-16	1.2e-18	1.1e-22	-
CRPSf						
	ECMWF	NCEP	ECCC	KMA	L_2 -bary	$GaussW_2$ -bary
L_2 -bary	1.4e-19	1.4e-26	6.3e-31	1.2e-26	-	2.1e-03
$GaussW_2$ -bary	2.3e-20	1.3e-26	6.0e-31	8.4e-27	2.1e-03	-

Table D3. Two-sided Wilcoxon test’s p-value for the geopotential height (z500) for weeks 3 and 4 combined. Bold indicates that the models are significantly different from each other at the 5% significance level (i.e. p-values< 0.05) with respect to the chosen score.

CRPSmean						
	ECMWF	NCEP	ECCC	KMA	L_2 -bary	$GaussW_2$ -bary
L_2 -bary	9.4e-03	8.1e-08	4.8e-04	3.7e-06	-	3.0e-01
$GaussW_2$ -bary	1.0e-02	7.2e-08	4.3e-04	3.2e-06	3.0e-01	-
CRPSp						
	ECMWF	NCEP	ECCC	KMA	L_2 -bary	$GaussW_2$ -bary
L_2 -bary	1.3e-03	3.6e-06	8.4e-05	1.1e-05	-	2.4e-12
$GaussW_2$ -bary	3.0e-05	9.4e-08	1.6e-06	1.2e-07	2.4e-12	-
CRPSf						
	ECMWF	NCEP	ECCC	KMA	L_2 -bary	$GaussW_2$ -bary
L_2 -bary	2.6e-09	2.0e-13	2.5e-17	6.2e-08	-	9.8e-01
$GaussW_2$ -bary		2.1e-09	1.9e-13	5.7e-17	6.5e-08	9.8e-01
						-

D2 Barycenters with three models

Table D4. Two-sided Wilcoxon test's p-value for the 2m temperature (t2m) for weeks 3 and 4 combined. Bold indicates that the models are significantly different from each other at the 5% significance level (i.e. p-values < 0.05).

CRPSmean								
	ECMWF	NCEP	ECCC	KMA	<i>L</i> ₂ -bary		<i>GaussW</i> ₂ -bary	
					4 models	3 models	4 models	3 models
<i>L</i> ₂ -bary (3 models)	6.4e-01	1.4e-07	4.4e-06	3.3e-07	4.1e-06	-	5.6e-06	1.8e-01
<i>GaussW</i> ₂ -bary (3 models)	5.8e-01	1.5e-07	5.7e-06	4.5e-07	6.8e-06	1.8e-01	2.7e-06	-
CRPSP								
	ECMWF	NCEP	ECCC	KMA	<i>L</i> ₂ -bary		<i>GaussW</i> ₂ -bary	
					4 models	3 models	4 models	3 models
<i>L</i> ₂ -bary (3 models)	8.9e-01	3.3e-08	1.4e-06	2.0e-10	3.6e-06	-	5.4e-12	3.4e-10
<i>GaussW</i> ₂ -bary (3 models)	2.7e-01	9.3e-11	1.5e-08	1.8e-12	2.1e-01	3.4e-10	8.3e-07	-
CRPSf								
	ECMWF	NCEP	ECCC	KMA	<i>L</i> ₂ -bary		<i>GaussW</i> ₂ -bary	
					4 models	3 models	4 models	3 models
<i>L</i> ₂ -bary (3 models)	1.2e-01	6.8e-13	5.4e-17	7.2e-16	1.5e-07	-	6.7e-10	4.1e-02
<i>GaussW</i> ₂ -bary (3 models)	7.6e-02	6.4e-14	5.3e-17	2.7e-16	9.7e-06	4.1e-02	2.5e-08	-

Table D5. Two-sided Wilcoxon test's p-value for the 10m wind speed (ws10m) for weeks 3 and 4 combined. Bold indicates that the models are significantly different from each other at the 5% significance level (i.e. p-values < 0.05).

CRPSmean								
	ECMWF	NCEP	ECCC	KMA	L_2 -bary		$GaussW_2$ -bary	
					4 models	3 models	4 models	3 models
L_2 -bary (3 models)	9.9e-01	4.9e-13	7.7e-09	3.3e-15	9.1e-10	-	4.6e-10	4.8e-01
$GaussW_2$ -bary (3 models)	8.2e-01	5.9e-13	9.2e-09	2.0e-15	2.1e-07	4.8e-01	2.9e-09	-
CRPSp								
	ECMWF	NCEP	ECCC	KMA	L_2 -bary		$GaussW_2$ -bary	
					4 models	3 models	4 models	3 models
L_2 -bary (3 models)	6.1e-01	3.1e-09	5.8e-08	1.3e-12	3.1e-06	-	9.7e-17	4.5e-20
$GaussW_2$ -bary (3 models)	5.1e-02	6.4e-14	1.6e-11	2.6e-17	8.9e-01	4.5e-20	1.1e-06	-
CRPSf								
	ECMWF	NCEP	ECCC	KMA	L_2 -bary		$GaussW_2$ -bary	
					4 models	3 models	4 models	3 models
L_2 -bary (3 models)	3.3e-09	6.9e-22	8.8e-30	1.4e-25	2.3e-11	-	4.2e-13	3.8e-02
$GaussW_2$ -bary (3 models)	4.1e-10	4.3e-22	6.2e-30	7.7e-26	1.7e-08	3.8e-02	2.1e-11	-

Table D6. Two-sided Wilcoxon test's p-value for the geopotential height (z500) for weeks 3 and 4 combined. Bold indicates that the models are significantly different from each other at the 5% significance level (i.e. p-values < 0.05).

CRPSmean								
	ECMWF	NCEP	ECCC	KMA	<i>L</i> ₂ -bary		<i>GaussW</i> ₂ -bary	
					4 models	3 models	4 models	3 models
<i>L</i> ₂ -bary (3 models)	2.0e-01	7.9e-07	3.3e-03	2.1e-05	1.3e-02	-	1.0e-02	2.4e-01
<i>GaussW</i> ₂ -bary (3 models)	1.9e-01	8.9e-07	2.8e-03	2.4e-05	2.0e-02	2.4e-01	9.2e-03	-
CRPSp								
	ECMWF	NCEP	ECCC	KMA	<i>L</i> ₂ -bary		<i>GaussW</i> ₂ -bary	
					4 models	3 models	4 models	3 models
<i>L</i> ₂ -bary (3 models)	7.2e-02	4.9e-06	6.5e-04	8.8e-05	7.9e-02	-	5.9e-05	2.8e-07
<i>GaussW</i> ₂ -bary (3 models)	1.8e-02	3.7e-07	4.3e-05	2.9e-06	5.7e-01	2.8e-07	3.1e-02	-
CRPSf								
	ECMWF	NCEP	ECCC	KMA	<i>L</i> ₂ -bary		<i>GaussW</i> ₂ -bary	
					4 models	3 models	4 models	3 models
<i>L</i> ₂ -bary (3 models)	4.3e-05	9.2e-13	1.0e-13	8.6e-06	5.4e-03	-	2.3e-03	8.3e-01
<i>GaussW</i> ₂ -bary (3 models)	2.5e-05	5.9e-13	5.5e-14	9.2e-06	1.3e-02	8.3e-01	3.2e-03	-

Author contributions. Conceptualization: CLC, AT, RF and RP; funding acquisition: AT and RF; formal analysis: CLC; methodology: CLC, AT, RF and RP; writing - original draft report: CLC; writing - review and editing: AT, RF and RP.

Competing interests. The authors declare that they have no conflict of interest.

565 *Acknowledgements.* The authors thank Naveen Goutham for code and advice on forecast calibration. The authors thank the institut de Mathématiques pour la Planète Terre for (partial) funding (iMPT-AAP2021). This work was partially supported by the grant ANR-23-ERCC-0006 from Agence nationale de la recherche (ANR). This work is supported by Hi! PARIS and ANR/France 2030 program (ANR-23-IACL-0005). This research was produced within the framework of Energy4Climate Interdisciplinary Center (E4C) of IP Paris and Ecole des Ponts ParisTech. This research was supported by 3rd Programme d’Investissements d’Avenir [ANR-18-EUR-0006-02].

570 References

- Agueh, M. and Carlier, G.: Barycenters in the Wasserstein Space, *SIAM Journal on Mathematical Analysis*, 43, 904–924, <https://doi.org/10.1137/100805741>, 2011.
- Alessandri, A., Borrelli, A., Navarra, A., Arribas, A., Déqué, M., Rogel, P., and Weisheimer, A.: Evaluation of Probabilistic Quality and Value of the ENSEMBLES Multimodel Seasonal Forecasts: Comparison with DEMETER, *Monthly Weather Review*, 139, 581 – 607, <https://doi.org/10.1175/2010MWR3417.1>, 2011.
- Backhoff-Veraguas, Julio, Fontbona, Joaquin, Rios, Gonzalo, and Tobar, Felipe: Bayesian learning with Wasserstein barycenters*, *ESAIM: PS*, 26, 436–472, <https://doi.org/10.1051/ps/2022015>, 2022.
- Becker, E., van den Dool, H., and Zhang, Q.: Predictability and Forecast Skill in NMME, *Journal of Climate*, 27, 5891 – 5906, <https://doi.org/10.1175/JCLI-D-13-00597.1>, 2014.
- 580 Bertino, L., Evensen, G., and Wackernagel, H.: Sequential Data Assimilation Techniques in Oceanography, *International Statistical Review*, 71, 223–241, <https://doi.org/https://doi.org/10.1111/j.1751-5823.2003.tb00194.x>, 2003.
- Casanova, S. and Ahrens, B.: On the Weighting of Multimodel Ensembles in Seasonal and Short-Range Weather Forecasting, *Monthly Weather Review*, 137, 3811 – 3822, <https://doi.org/10.1175/2009MWR2893.1>, 2009.
- DelSole, T., Yang, X., and Tippett, M. K.: Is unequal weighting significantly better than equal weighting for multi-model forecasting?, *Quarterly Journal of the Royal Meteorological Society*, 139, 176–183, <https://doi.org/10.1002/qj.1961>, 2013.
- 585 Eade, R., Smith, D., Scaife, A., Wallace, E., Dunstone, N., Hermanson, L., and Robinson, N.: Do seasonal-to-decadal climate predictions underestimate the predictability of the real world?, *Geophysical Research Letters*, 41, 5620–5628, <https://doi.org/https://doi.org/10.1002/2014GL061146>, 2014.
- Ferrone, A., Mastrangelo, D., and Malguzzi, P.: Multimodel probabilistic prediction of 2 m-temperature anomalies on the monthly timescale, *Advances in Science and Research*, 14, 123–129, <https://doi.org/10.5194/asr-14-123-2017>, 2017.
- 590 Flamary, R., Lounici, K., and Ferrari, A.: Concentration bounds for linear Monge mapping estimation and optimal transport domain adaptation, <https://arxiv.org/abs/1905.10155>, 2020.
- Fortin, V., Abaza, M., Anctil, F., and Turcotte, R.: Why Should Ensemble Spread Match the RMSE of the Ensemble Mean?, *Journal of Hydrometeorology*, 15, 1708 – 1713, <https://doi.org/10.1175/JHM-D-14-0008.1>, 2014.
- 595 Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., da Silva, A. M., Gu, W., Kim, G.-K., Koster, R., Lucchesi, R., Merkova, D., Nielsen, J. E., Partyka, G., Pawson, S., Putman, W., Rienecker, M., Schubert, S. D., Sienkiewicz, M., and Zhao, B.: The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2), *Journal of Climate*, 30, 5419 – 5454, <https://doi.org/10.1175/JCLI-D-16-0758.1>, 2017.
- 600 Global Modeling and Assimilation Office (GMAO): MERRA-2 statD_2d_slv_Nx: 2d, Daily,Aggregated Statistics,Single-Level,Assimilation,Single-Level Diagnostics V5.12.4, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC), <https://doi.org/10.5067/9SC1VNTWGWV3>, Last accessed 18/09/2024, 2015a.
- Global Modeling and Assimilation Office (GMAO): MERRA-2 inst1_2d_asm_Nx: 2d,1-Hourly,Instantaneous,Single-Level,Assimilation,Single-Level Diagnostics V5.12.4, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC), <https://doi.org/10.5067/3Z173KIE2TPD>, Last accessed 19/09/2024, 2015b.
- 605

- Global Modeling and Assimilation Office (GMAO): MERRA-2 inst3_3d_asm_Np: 3d,3-Hourly,Instantaneous,Pressure-Level,Assimilation,Assimilated Meteorological Fields V5.12.4, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC), <https://doi.org/10.5067/QBZ6MG944HW0>, Last accessed 19/09/2024, 2015c.
- Gnassounou, T., Flamary, R., and Gramfort, A.: Convolution Monge Mapping Normalization for learning on sleep data, in: Advances in Neural Information Processing Systems, edited by Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., vol. 36, pp. 10457–10476, Curran Associates, Inc., https://proceedings.neurips.cc/paper_files/paper/2023/file/21718991f6acf19a42376b5c7a8668c5-Paper-Conference.pdf, 2023.
- Gneiting, T. and Raftery, A. E.: Strictly Proper Scoring Rules, Prediction, and Estimation, *Journal of the American Statistical Association*, 102, 359–378, <https://doi.org/10.1198/016214506000001437>, 2007.
- Gneiting, T., Raftery, A. E., Westveld, A. H., and Goldman, T.: Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation, *Monthly Weather Review*, 133, 1098 – 1118, <https://doi.org/10.1175/MWR2904.1>, 2005.
- Gonzalez, P. L. M., Brayshaw, D. J., and Ziel, F.: A new approach to extended-range multimodel forecasting: Sequential learning algorithms, *Quarterly Journal of the Royal Meteorological Society*, 147, 4269–4282, <https://doi.org/10.1002/qj.4177>, 2021.
- Goutham, N., Plougonven, R., Omrani, H., Parey, S., Tankov, P., Tantet, A., Hitchcock, P., and Drobinski, P.: How Skillful Are the European Subseasonal Predictions of Wind Speed and Surface Temperature?, *Monthly Weather Review*, 150, 1621 – 1637, <https://doi.org/10.1175/MWR-D-21-0207.1>, 2022.
- Hagedorn, R., Doblas-Reyes, F. J., and Palmer, T.: The rationale behind the success of multi-model ensembles in seasonal forecasting — I. Basic concept, *Tellus A: Dynamic Meteorology and Oceanography*, 57, 219–233, <https://doi.org/10.3402/tellusa.v57i3.14657>, 2005.
- Hagedorn, R., Buizza, R., Hamill, T. M., Leutbecher, M., and Palmer, T. N.: Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts, *Quarterly Journal of the Royal Meteorological Society*, 138, 1814–1827, <https://doi.org/10.1002/qj.1895>, 2012.
- Hamill, T. M.: Verification of TIGGE Multimodel and ECMWF Reforecast-Calibrated Probabilistic Precipitation Forecasts over the Contiguous United States, *Monthly Weather Review*, 140, 2232 – 2252, <https://doi.org/10.1175/MWR-D-11-00220.1>, 2012.
- Haughton, N., Abramowitz, G., Pitman, A., and Phipps, S. J.: Weighting climate model ensembles for mean and variance estimates, *Climate Dynamics*, 45, 3169–3181, <https://doi.org/10.1007/s00382-015-2531-3>, 2015.
- Heizenreder, D., Trepte, S., and Denhard, M.: SRNWP-PEPS: A regional multi-model ensemble in Europe, *The European Forecaster: Newsletter of the WGCEF*, 11, 2006.
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, *Quarterly Journal of the Royal Meteorological Society*, 146, 1999–2049, <https://doi.org/10.1002/qj.3803>, 2020.
- Kalnay, E.: Atmospheric modeling, data assimilation and predictability, Cambridge University Press, 2003.
- Karpechko, A. Y., Charlton-Perez, A., Balmaseda, M., Tyrrell, N., and Vitart, F.: Predicting Sudden Stratospheric Warming 2018 and Its Climate Impacts With a Multimodel Ensemble, *Geophysical Research Letters*, 45, 13,538–13,546, <https://doi.org/10.1029/2018GL081091>, 2018.

- Kharin, V. V. and Zwiers, F. W.: Climate Predictions with Multimodel Ensembles, *Journal of Climate*, 15, 793 – 799, [https://doi.org/10.1175/1520-0442\(2002\)015<0793:CPWME>2.0.CO;2](https://doi.org/10.1175/1520-0442(2002)015<0793:CPWME>2.0.CO;2), 2002.
- 645 Kioutsoukakis, I. and Galmarini, S.: De praeceptis ferendis: good practice in multi-model ensembles, *Atmospheric Chemistry and Physics*, 14, 11 791–11 815, <https://doi.org/10.5194/acp-14-11791-2014>, 2014.
- Kirtman, B. P., Min, D., Infanti, J. M., Kinter, J. L., Paolino, D. A., Zhang, Q., van den Dool, H., Saha, S., Mendez, M. P., Becker, E., Peng, P., Tripp, P., Huang, J., DeWitt, D. G., Tippett, M. K., Barnston, A. G., Li, S., Rosati, A., Schubert, S. D., Rienecker, M., Suarez, M., Li, Z. E., Marshak, J., Lim, Y.-K., Tribbia, J., Pegion, K., Merryfield, W. J., Denis, B., and Wood, E. F.: The North American Multimodel
- 650 Ensemble: Phase-1 Seasonal-to-Interannual Prediction; Phase-2 toward Developing Intraseasonal Prediction, *Bulletin of the American Meteorological Society*, 95, 585 – 601, <https://doi.org/10.1175/BAMS-D-12-00050.1>, 2014.
- Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., and Eyring, V.: A climate model projection weighting scheme accounting for performance and interdependence, *Geophysical Research Letters*, 44, 1909–1918, <https://doi.org/https://doi.org/10.1002/2016GL072012>, 2017.
- 655 Le Coz, C., Tantet, A., Flamary, R., and Plougonven, R.: Code for "A barycenter-based approach for the multi-model ensembling of subseasonal forecasts", <https://doi.org/10.5281/zenodo.15058503>, 2025a.
- Le Coz, C., Tantet, A., Flamary, R., and Plougonven, R.: Pre-processed data for "A barycenter-based approach for the multi-model ensembling of subseasonal forecasts", <https://doi.org/10.5281/zenodo.15038871>, 2025b.
- Leung, L. R., Hamlet, A. F., Lettenmaier, D. P., and Kumar, A.: Simulations of the ENSO Hydroclimate Signals in the Pacific
- 660 Northwest Columbia River Basin, *Bulletin of the American Meteorological Society*, 80, 2313 – 2330, [https://doi.org/10.1175/1520-0477\(1999\)080<2313:SOTEHS>2.0.CO;2](https://doi.org/10.1175/1520-0477(1999)080<2313:SOTEHS>2.0.CO;2), 1999.
- Lussana, C., Nipen, T. N., Seierstad, I. A., and Elo, C. A.: Ensemble-based statistical interpolation with Gaussian anamorphosis for the spatial analysis of precipitation, *Nonlinear Processes in Geophysics*, 28, 61–91, <https://doi.org/10.5194/npg-28-61-2021>, 2021.
- Manrique-Suñén, A., Gonzalez-Reviriego, N., Torralba, V., Cortesi, N., and Doblas-Reyes, F. J.: Choices in the Verification of S2S Forecasts
- 665 and Their Implications for Climate Services, *Monthly Weather Review*, 148, 3995 – 4008, <https://doi.org/10.1175/MWR-D-20-0067.1>, 2020.
- Manzanas, R., Gutiérrez, J. M., Bhend, J., Hemri, S., Doblas-Reyes, F. J., Torralba, V., Penabad, E., and Brookshaw, A.: Bias adjustment and ensemble recalibration methods for seasonal forecasting: a comprehensive intercomparison using the C3S dataset, *Climate Dynamics*, 53, 1287–1305, <https://doi.org/10.1007/s00382-019-04640-4>, 2019.
- 670 Materia, S., Ángel G. Muñoz, Álvarez Castro, M. C., Mason, S. J., Vitart, F., and Gualdi, S.: Multimodel Subseasonal Forecasts of Spring Cold Spells: Potential Value for the Hazelnut Agribusiness, *Weather and Forecasting*, 35, 237 – 254, <https://doi.org/10.1175/WAF-D-19-0086.1>, 2020.
- Matheson, J. E. and Winkler, R. L.: Scoring Rules for Continuous Probability Distributions, *Management Science*, 22, 1087–1096, <http://www.jstor.org/stable/2629907>, 1976.
- 675 Ning, L., Carli, F. P., Ebtehaj, A. M., Foufoula-Georgiou, E., and Georgiou, T. T.: Coping with model error in variational data assimilation using optimal mass transport, *Water Resources Research*, 50, 5817–5830, <https://doi.org/10.1002/2013WR014966>, 2014.
- Palmer, T. N., Alessandri, A., Andersen, U., Cantelaube, P., Davey, M., Décluse, P., Déqué, M., Díez, E., Doblas-Reyes, F. J., Feddersen, H., Graham, R., Gualdi, S., Guérémy, J.-F., Hagedorn, R., Hoshen, M., Keenlyside, N., Latif, M., Lazar, A., Maisonnave, E., Marletto, V., Morse, A. P., Orfila, B., Rogel, P., Terres, J.-M., and Thomson, M. C.: Development of a European Multimodel En-

- semble System for Seasonal-to-Interannual Prediction (DEMETER), *Bulletin of the American Meteorological Society*, 85, 853 – 872, <https://doi.org/10.1175/BAMS-85-6-853>, 2004.
- Papayiannis, G. I., Galanis, G. N., and Yannacopoulos, A. N.: Model aggregation using optimal transport and applications in wind speed forecasting, *Environmetrics*, 29, e2531, <https://doi.org/10.1002/env.2531>, 2018.
- Pegion, K., Kirtman, B. P., Becker, E., Collins, D. C., LaJoie, E., Burgman, R., Bell, R., DelSole, T., Min, D., Zhu, Y., Li, W., Sinsky, E., Guan, H., Gottschalck, J., Metzger, E. J., Barton, N. P., Achuthavarier, D., Marshak, J., Koster, R. D., Lin, H., Gagnon, N., Bell, M., Tippett, M. K., Robertson, A. W., Sun, S., Benjamin, S. G., Green, B. W., Bleck, R., and Kim, H.: The Subseasonal Experiment (SubX): A Multimodel Subseasonal Prediction Experiment, *Bulletin of the American Meteorological Society*, 100, 2043 – 2060, <https://doi.org/10.1175/BAMS-D-18-0270.1>, 2019.
- Peyré, G. and Cuturi, M.: *Computational Optimal Transport*, 2020.
- Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M.: Using Bayesian Model Averaging to Calibrate Forecast Ensembles, *Monthly Weather Review*, 133, 1155 – 1174, <https://doi.org/10.1175/MWR2906.1>, 2005.
- Rajagopalan, B., Lall, U., and Zebiak, S. E.: Categorical Climate Forecasts through Regularization and Optimal Combination of Multiple GCM Ensembles, *Monthly Weather Review*, 130, 1792 – 1811, [https://doi.org/10.1175/1520-0493\(2002\)130<1792:CCFTRA>2.0.CO;2](https://doi.org/10.1175/1520-0493(2002)130<1792:CCFTRA>2.0.CO;2), 2002.
- Robertson, A. W., Lall, U., Zebiak, S. E., and Goddard, L.: Improved Combination of Multiple Atmospheric GCM Ensembles for Seasonal Prediction, *Monthly Weather Review*, 132, 2732 – 2744, <https://doi.org/10.1175/MWR2818.1>, 2004.
- Robin, Y., Yiou, P., and Naveau, P.: Detecting changes in forced climate attractors with Wasserstein distance, *Nonlinear Processes in Geophysics*, 24, 393–405, <https://doi.org/10.5194/npg-24-393-2017>, 2017.
- Robin, Y., Vrac, M., Naveau, P., and Yiou, P.: Multivariate stochastic bias corrections with optimal transport, *Hydrology and Earth System Sciences*, 23, 773–786, <https://doi.org/10.5194/hess-23-773-2019>, 2019.
- Santambrogio, F.: Progress in Nonlinear Differential Equations and Their Applications, in: *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, vol. 87, Birkhäuser Cham, <https://doi.org/10.1007/978-3-319-20828-2>, 2015.
- Schulzweida, U.: CDO User Guide, <https://doi.org/10.5281/zenodo.7112925>, 2022.
- Smith, D. M., Scaife, A. A., Boer, G. J., Caian, M., Doblas-Reyes, F. J., Guemas, V., Hawkins, E., Hazeleger, W., Hermanson, L., Ho, C. K., Ishii, M., Kharin, V., Kimoto, M., Kirtman, B., Lean, J., Matei, D., Merryfield, W. J., Müller, W. A., Pohlmann, H., Rosati, A., Wouters, B., and Wyser, K.: Real-time multi-model decadal climate predictions, *Climate Dynamics*, 41, 2875–2888, <https://doi.org/10.1007/s00382-012-1600-0>, 2013.
- Specq, D., Batté, L., Déqué, M., and Ardilouze, C.: Multimodel Forecasting of Precipitation at Subseasonal Timescales Over the Southwest Tropical Pacific, *Earth and Space Science*, 7, e2019EA001 003, <https://doi.org/10.1029/2019EA001003>, 2020.
- Takaya, Y.: Chapter 12 - Forecast System Design, Configuration, and Complexity, in: *Sub-Seasonal to Seasonal Prediction*, edited by Robertson, A. W. and Vitart, F., pp. 245–259, Elsevier, <https://doi.org/10.1016/B978-0-12-811714-9.00012-7>, 2019.
- Vigaud, N., Robertson, A. W., and Tippett, M. K.: Multimodel Ensembling of Subseasonal Precipitation Forecasts over North America, *Monthly Weather Review*, 145, 3913 – 3928, <https://doi.org/10.1175/MWR-D-17-0092.1>, 2017.
- Vigaud, N., Tippett, M. K., Yuan, J., Robertson, A. W., and Acharya, N.: Spatial Correction of Multimodel Ensemble Subseasonal Precipitation Forecasts over North America Using Local Laplacian Eigenfunctions, *Monthly Weather Review*, 148, 523 – 539, <https://doi.org/10.1175/MWR-D-19-0134.1>, 2020.
- Villani, C.: *Topics in Optimal Transportation*, American Mathematical Society, 2003.

- Vissio, G. and Lucarini, V.: Evaluating a stochastic parametrization for a fast–slow system using the Wasserstein distance, *Nonlinear Processes in Geophysics*, 25, 413–427, <https://doi.org/10.5194/npg-25-413-2018>, 2018.
- 720 Vissio, G., Lembo, V., Lucarini, V., and Ghil, M.: Evaluating the Performance of Climate Models Based on Wasserstein Distance, *Geophysical Research Letters*, 47, e2020GL089385, <https://doi.org/10.1029/2020GL089385>, 2020.
- Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Déqué, M., Ferranti, L., Fucile, E., Fuentes, M., Hendon, H., Hodgson, J., Kang, H.-S., Kumar, A., Lin, H., Liu, G., Liu, X., Malguzzi, P., Mallas, I., Manoussakis, M., Mastrangelo, D., MacLachlan, C., McLean, P., Minami, A., Mladek, R., Nakazawa, T., Najm, S., Nie, Y., Rixen, M., Robertson, A. W., Ruti, P., Sun, C., Takaya, Y., Tolstykh, M., Venuti, F., Waliser, D., Woolnough, S., Wu, T., Won, D.-J., Xiao, H., Zaripov, R., and Zhang, L.: The Subseasonal to Seasonal (S2S) Prediction Project Database, *Bulletin of the American Meteorological Society*, 98, 163 – 173, <https://doi.org/10.1175/BAMS-D-16-0017.1>, 2017.
- Wanders, N. and Wood, E. F.: Improved sub-seasonal meteorological forecast skill using weighted multi-model ensemble simulations, *Environmental Research Letters*, 11, 094007, <https://doi.org/10.1088/1748-9326/11/9/094007>, 2016.
- 730 Wang, Y., Ren, H.-L., Zhou, F., Fu, J.-X., Chen, Q.-L., Wu, J., Jie, W.-H., and Zhang, P.-Q.: Multi-Model Ensemble Sub-Seasonal Forecasting of Precipitation over the Maritime Continent in Boreal Summer, *Atmosphere*, 11, <https://doi.org/10.3390/atmos11050515>, 2020.
- Weigel, A. P., Liniger, M. A., and Appenzeller, C.: Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts?, *Quarterly Journal of the Royal Meteorological Society*, 134, 241–260, <https://doi.org/10.1002/qj.210>, 2008.
- Wilcoxon, F.: Individual Comparisons by Ranking Methods, *Biometrics Bulletin*, 1, 80–83, <http://www.jstor.org/stable/3001968>, 1945.
- 735 Wilks, D. S.: Chapter 9 - Forecast Verification, in: *Statistical Methods in the Atmospheric Sciences (Fourth Edition)*, edited by Wilks, D. S., pp. 369–483, Elsevier, fourth edition edn., ISBN 978-0-12-815823-4, <https://doi.org/10.1016/B978-0-12-815823-4.00009-2>, 2019.
- Zheng, C., Chang, E. K.-M., Kim, H., Zhang, M., and Wang, W.: Subseasonal to Seasonal Prediction of Wintertime Northern Hemisphere Extratropical Cyclone Activity by S2S and NMME Models, *Journal of Geophysical Research: Atmospheres*, 124, 12057–12077, <https://doi.org/10.1029/2019JD031252>, 2019.
- 740 Zhou, X., Zhu, Y., Hou, D., Fu, B., Li, W., Guan, H., Sinsky, E., Kolczynski, W., Xue, X., Luo, Y., Peng, J., Yang, B., Tallapragada, V., and Pegion, P.: The Development of the NCEP Global Ensemble Forecast System Version 12, *Weather and Forecasting*, 37, 1069 – 1084, <https://doi.org/10.1175/WAF-D-21-0112.1>, 2022.