Enhancing Seasonal Forecast Skills by Optimally Weighting the Ensemble from Fresh Data

JULIEN BRAJARD,[a] FRANÇOIS COUNILLON,[a,b] YIGUO WANG,[a] AND MADLEN KIMMRITZ[c]

[a] *Nansen Environmental and Remote Sensing Center, Bergen, Norway*
[b] *University of Bergen, Bergen Norway*
[c] *Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Bremerhaven , Germany*

(Manuscript received 28 September 2022, in final form 9 March 2023, accepted 28 April 2023)

ABSTRACT: Dynamical climate predictions are produced by assimilating observations and running ensemble simulations of Earth system models. This process is time consuming and by the time the forecast is delivered, new observations are already available, making it obsolete from the release date. Moreover, producing such predictions is computationally demanding, and their production frequency is restricted. We tested the potential of a computationally cheap weighting average technique that can continuously adjust such probabilistic forecasts—in between production intervals—using newly available data. The method estimates local positive weights computed with a Bayesian framework, favoring members closer to observations. We tested the approach with the Norwegian Climate Prediction Model (NorCPM), which assimilates monthly sea surface temperature (SST) and hydrographic profiles with the ensemble Kalman filter. By the time the NorCPM forecast is delivered operationally, a week of unused SST data are available. We demonstrate the benefit of our weighting method on retrospective hindcasts. The weighting method greatly enhanced the NorCPM hindcast skill compared to the standard equal weight approach up to a 2-month lead time (global correlation of 0.71 vs 0.55 at a 1-month lead time and 0.51 vs 0.45 at a 2-month lead time). The skill at a 1-month lead time is comparable to the accuracy of the EnKF analysis. We also show that weights determined using SST data can be used to improve the skill of other quantities, such as the sea ice extent. Our approach can provide a continuous forecast between the intermittent forecast production cycle and be extended to other independent datasets.

KEYWORDS: Ensembles; Forecasting techniques; Seasonal forecasting; Data assimilation; Postprocessing

1. Introduction

Climate prediction systems have become essential tools for climate services and can help mitigate the risks and identify potential opportunities due to the changing climate (Hewitt and Lowe 2018; Mariotti et al. 2020; Goutham et al. 2022). Subseasonal-to-seasonal (S2S) predictions (Vitart et al. 2017; Vitart and Robertson 2018; Becker et al. 2022)—an activity developed by the World Weather Research Programme/World Climate Research Programme S2S Prediction Project—provide predictions from 14 days up to 2 months, and seasonal to interannual predictions—e.g., ECMWF (SEAS5), Copernicus Climate change Service (C3S), North American Multimodel Ensemble (NMME) (Kirtman et al. 2014), and NorCPM (Wang et al. 2019)—provide forecasts from from lead month 1 up to a year (we refer to Meehl et al. 2021 for a detailed specification of each initiative). For such a time scale, deterministic predictions cannot be skillful because of the chaotic nature of the atmosphere (Palmer et al. 2014; Zhang et al. 2019). However, weather events can be modulated by interaction with the slower Earth system components (ocean, land, and sea ice), for which variability can be predicted. Such predictions are provided as ensemble forecasts that represent the forecast uncertainty. Most of the time, it assumed that all the members of the ensemble have the same likelihood, but it was shown that some properties of the forecast can be used to favor the more likely members (Thorey et al. 2017; Dobrynin et al. 2018).

The production of climate predictions is computationally expensive and time-demanding. It includes the collection of observations, their assimilation into a numerical model (Carrassi et al. 2018), the production of the ensemble forecast simulations, and a security buffer period to ensure timely delivery. Although increasing the production frequency of such forecasts can enhance their accuracy, it is limited for practical reasons, typically every month as for the European Centre for Medium-Range Weather Forecasts (ECMWF) (SEAS5), C3S, the North American Multimodel ensemble (Kirtman et al. 2014), and NorCPM (Wang et al. 2019) for seasonal forecasting systems. Newly available observations are becoming available during the forecast production, making the forecasts suboptimal from the moment they are produced. For example, forecasts are effectively available on the 7th day of the month for NorCPM, on the 9th day of the month for NMME on the 13th day of the month for C3S, and on the 5th day of the month for SEAS5, which implies that about 1–2 weeks of newly available data has been available since the assimilation step. Here, we propose a novel methodology to update the seasonal ensemble forecast based on new observations. The potential of such an approach is nicely exemplified by Lean et al. (2021), who showed that continuous integration of new

Denotes content that is immediately available upon publication as open access.

Corresponding author: Julien Brajard, julien.brajard@nersc.no

DOI: 10.1175/WAF-D-22-0166.1

observations during the assimilation procedure could lead to an increase of 2–3 h of additional predictive skill in the ECMWF numerical weather prediction forecast.

In the new method introduced, we test the benefit of adjusting the weights (likelihood) for each member—based on the observations—and the new ensemble forecast is provided as a weighted mean of the former, reducing its uncertainty. The update step is computationally cheap and fast as the dynamical members do not need to be rerun. The forecast can be continuously updated once new observations are made available.

We test the method with the Norwegian Climate Prediction Model (NorCPM), which can provide skillful forecasts for up to 12 lead months in several regions (Wang et al. 2019). The system combines the Norwegian Earth System Model (Bentsen et al. 2013) with the ensemble Kalman filter (Evensen 2003) and produces monthly operational forecast (see https://klimavarsling.no). The version we use assimilates SST and hydrographic profiles in the ocean component of the Earth system model with 60 members. Here, we demonstrate the potential of our method on existing seasonal hindcasts (retrospective predictions) that match the setting of the operational system. Hindcasts are started four times per year during 1985–2010, and each hindcast runs 60 realizations (ensemble members) for 13 months initialized from the EnKF reanalysis. With an EnKF, all members are equally likely. Consequently, the most likely forecast is the ensemble mean, and the ensemble spreads provide a quantification of the uncertainty of the forecast (Evensen 2003). We test the benefit of adjusting the weights (likelihood) for each member based on the new SST observations. We focus on SST because it is available in near real time (with a lag of 1 day) and because it has been shown sufficient to constrain the variability in many regions of the Earth system, particularly in the tropics (Shukla 1998; Zhu et al. 2017; Wang et al. 2019). On the contrary, the hydrographic profiles also assimilated in NorCPM are available (in preliminary mode, meaning without final quality control) with a latency of 1-month delays. However, our method can be extended to any other observations.

2. Data

In the following, unless stated otherwise, t_k denotes the time of the variables with a regular monthly time period, i.e., $t_{k+1} - t_k = 1$ month.

a. Observations

We use weekly SST data from the National and Atmospheric Administration (NOAA) Optimum Interpolation SST (OISST) version 2 (Reynolds et al. 2002), which is made available by NOAA with one day of delay. SST are gridded at a 1° resolution—the entire field can be stacked into a vector of size $p = 39\,080$. The monthly SST ($\mathbf{y}_k \in \mathbb{R}^p$) at t_k is computed by averaging the weekly SST observations available during the month. For the sake of simplicity, we have considered that a week belongs to a month when the starting day is within the month. The observations are defined as the anomalies from the monthly climatology computed from 1982 to 2010. For simplicity, we refer to SST anomalies observations as "SST observations" in the following.

For each grid point, OISST also provides an error estimate. The observation error variance vector ($\boldsymbol{\sigma}_k^2 \in \mathbb{R}^p$) is the average of all the weekly variance error fields. Consistently with the NorCPM setting (e.g., Wang et al. 2019), we assume that observation errors are uncorrelated, i.e., the variance–covariance observation error matrix is diagonal—$\mathbf{R}_k \in \mathbb{R}^{p\times p}$ as

$$\mathbf{R}_k = \boldsymbol{\sigma}_k^2 \mathbf{I}_p, \tag{1}$$

where $\mathbf{I}_p$ is the identity matrix of size p.

b. Reanalyses and hindcasts

We use a reanalysis and hindcast dataset from NorCPM, which combines the Norwegian Earth System Model (Bentsen et al. 2013) and an ensemble Kalman filter (Evensen 2003). This system version is comparable to the one providing operational forecast, namely, it assimilates sea surface temperature and hydrographic profiles (temperature and salinity) using 60 members and strongly coupled data assimilation between the ocean and sea ice component (Bethke et al. 2021)—meaning that the ocean data also correct the sea ice component. We perform anomaly assimilation, meaning that the climatological monthly mean of the observations and the model are removed before comparing the two. The monthly climatology of the model is constructed from the 60-member historical ensemble run (without assimilation) over the period 1982–2010. For the hydrographic profiles, it is constructed from EN4 objective analysis (Good et al. 2013). Only the ocean and sea ice are directly updated by the data assimilation. The other components of the model (atmosphere, land) are adjusting dynamically through the coupling in between the monthly assimilation steps. The initial ensemble at the start of the reanalysis in 1980 is constructed by selecting 60 random initial conditions from a stable preindustrial simulation and integrating the ensemble from 1850 to 1980 using historical forcings from the Coupled Model Intercomparison Project version 5 (Taylor et al. 2012).

The seasonal hindcasts start on 15 January, 15 April, 15 July, and 15 October each year from 1985 to 2010, i.e., in total 104 hindcasts (26 years with four hindcasts per year). This is, therefore, less frequent than our operational forecast that is produced every month, but producing such a dataset is computational and storage demanding. In the present work, we use a monthly time sampling by default: for example, if the start date is on 15 January, the reanalysis month (or lead month 0) is the month of January from 1 to 31 January and observations from the whole month of January are used (as a matter of fact we assimilate the monthly average product from NOAA). Following this, lead month 1 corresponds to February. Each hindcast runs 60 realizations (ensemble members) for 13 months, initialized from the corresponding member in the reanalysis. In the case of a start in January, we estimate that by the time the process of assimilating observations for January month and of producing the 60 hindcasts is done, SST observations from the first week of February are available.

We start by analyzing the capability of our method to predict anomalies of SST. We also assess the performance for sea ice extent (see section 4d). For each starting month t_k (January, April, July, and October) between 1985 and 2010 and each member n ($1 \le n \le N = 60$), the anomaly of the SST at lead

month 0 is denoted $\mathbf{x}^f_{k,0,n}$ and corresponds to the analysis. The corresponding hindcast for the lead month h is denoted $\mathbf{x}^f_{k,h,n}$. The maximum hindcast lead month considered is 5 ($H = 5$).

At the date t_k, we have the following model states:

$$\mathbf{X}_{k,n} = \{\mathbf{x}^f_{k,0,n}, \mathbf{x}^f_{k,1,n}, \ldots, \mathbf{x}^f_{k,H,n}\}. \tag{2}$$

3. Optimal weight (OW) method

a. Main assumptions

To improve a given hindcast, we make use of SST observations in the first week of the hindcast. We compute a monthly average of SST observations shifted by one week ($\mathbf{y}_{k+1w}$).

Ideally, we would have liked to base the method on weekly average output, but only monthly outputs were saved (because of storage limitation) with starting months between 1985 and 2010. We use the following approximation to estimate the SST monthly hindcast shifted by one week:

$$\mathbf{x}_{k+1w,n} = \frac{3}{4}\mathbf{x}^f_{k,0,n} + \frac{1}{4}\mathbf{x}^f_{k,1,n}. \tag{3}$$

Two important assumptions are made. First, we assume that the EnKF with a finite size ensemble can approximate the system state's probability density function (pdf) as

$$f_{\text{pdf}}(\mathbf{X}_k) = \frac{1}{N}\sum_{n=1}^{N} \delta(\mathbf{X}_k - \mathbf{X}_{k,n}), \tag{4}$$

where f_{pdf} denotes the pdf, $\mathbf{X}_k$ is the random variable of the state with the nth ensemble member $\mathbf{X}_{k,n}$ defined as in Eq. (2), and δ is the Dirac delta measure. The expectation of the state $\mathbf{X}_k$ can thus be expressed by the arithmetic average of the ensemble and corresponds to the equal weights estimation:

$$\mathbb{E}(\mathbf{X}_k) = \frac{1}{N}\sum_{n=1}^{N} \mathbf{X}_{k,n}. \tag{5}$$

This result implies that the "equal weights" hindcast (typically used in seasonal forecasting and hereafter referred to as EW) is the optimal estimate without other sources of information. The EW will serve as a benchmark and is defined as follows:

$$\mathbf{x}^{\text{EW}}_{k,h} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}^f_{k,h,n}. \tag{6}$$

Second, as it is done in NorCPM (Wang et al. 2019), we assume the likelihood $f_{\text{pdf}}(\mathbf{y}_{k+1w}/\mathbf{X}_k)$ is Gaussian, defined as

$$f_{\text{pdf}}(\mathbf{y}_{k+1w}|\mathbf{X}_{k,n}) \propto \exp\left(-\frac{1}{2}\mathbf{d}^{\text{T}}_{k,n}\mathbf{R}^{-1}_k\mathbf{d}_{k,n}\right),$$

where $\mathbf{d}_{k,n} = \mathbf{y}_{k+1w} - \mathbf{x}_{k+1w,n}$ is the innovation, and $\mathbf{R}_k$ is the error-covariance observation error defined in Eq. (1)

b. Weight determination

In the optimal weight method (hereafter referred to as OW), the weights are determined as in the particle filter (see e.g., van Leeuwen et al. 2019; Evensen et al. 2022, for more details). The a posteriori density function can be formulated as follows with Bayes's formula:

$$\begin{aligned} f_{\text{pdf}}(\mathbf{X}_k|\mathbf{y}_{k+1w}) &= \frac{f_{\text{pdf}}(\mathbf{y}_{k+1w}|\mathbf{X}_k)}{f_{\text{pdf}}(\mathbf{y}_{k+1w})} f_{\text{pdf}}(\mathbf{X}_k) \\ &= \frac{1}{N}\sum_{n=1}^{N} \frac{f_{\text{pdf}}(\mathbf{y}_{k+1w}|\mathbf{X}_k)}{f_{\text{pdf}}(\mathbf{y}_{k+1w})}\delta(\mathbf{X}_k - \mathbf{X}_{k,n}) \\ &= \frac{1}{N f_{\text{pdf}}(\mathbf{y}_{k+1w})}\sum_{n=1}^{N} f_{\text{pdf}}(\mathbf{y}_{k+1w}|\mathbf{X}_{k,n})\delta(\mathbf{X}_k - \mathbf{X}_{k,n}) \\ &\propto \sum_{n=1}^{N} \exp\left(-\frac{1}{2}\mathbf{d}^{\text{T}}_{k,n}\mathbf{R}^{-1}_k\mathbf{d}_{k,n}\right)\delta(\mathbf{X}_k - \mathbf{X}_{k,n}). \end{aligned} \tag{8}$$

We can then define the weighted pdf as

$$f_{\text{pdf}}(\mathbf{X}_k|\mathbf{y}_{k+1w}) = \sum_{n=1}^{N} w_{k,n}\delta(\mathbf{X}_k - \mathbf{X}_{k,n}), \tag{9}$$

where the weights $w_{k,n} \propto \exp[-(1/2)\mathbf{d}^{\text{T}}_{k,n}\mathbf{R}^{-1}_k\mathbf{d}_{k,n}]$. The weights are positive, and their sum is set equal to 1. They are optimal in the sense that they maximize the a posteriori density function. Since, in this case, the weights would not vary in space, they are called "global."

We have $p = 39\,080$ observations and there is a degeneracy of the particle filter for big systems—called the "curse of dimensionality" (Snyder et al. 2008). It means that estimating weights globally as in Eq. (9) results in having all weights close to zeros except for one. It is equivalent to selecting only one member of the ensemble, which strongly alters the reliability of the ensemble forecast. Therefore, the global weight approach can only be applied if the dimension of the system is sufficiently small to avoid this problem, which is not the case here.

Hence, in our case, we estimate localized weights computed at each grid point from local observations. To do that, we restrict the innovation locally around the grid point of interest i ($0 < i < p$):

$$\mathbf{d}^{[i]}_{k,n} = \boldsymbol{\rho}_i \circ \mathbf{d}_{k,n}, \tag{10}$$

where $\circ$ is the Schür product and $\boldsymbol{\rho}_i$ is a tapering vector whom element j is defined as

$$\rho_{ij} = f[d(i, j)/L], \tag{11}$$

where $d(i, j)$ is the distance between grid points i and j, and L is the localization radius. The tapering function f is the Gaspari–Cohn function (Gaspari and Cohn 1999), which decreases from 1 at the target point [$d(i, j) = 0$] to 0 beyond the localization radius [$d(i, j) > L$]. It ensures continuity in the estimated weights. The localization reduces the effective number of observations used to determine the weights and thus mitigates the degeneracy described above.

Inflating the observation error by a multiplicative factor λ° is also used to counteract particle filter degeneracy. If $\lambda^\circ = 1$, the inflation has no effect, while if $\lambda^\circ = \infty$, the observations will have no influence, and the weight estimate converges to EW.

When combining localization and inflation, the final expression of the weight for an initial date t_k, a member n, and a grid point i is

$$w_{k,n,i} \propto \exp\left\{-\frac{1}{2}(\boldsymbol{\rho}_i \circ \mathbf{d}_{k,n})^{\mathrm{T}}[(\lambda^{\circ})^2\mathbf{R}]^{-1}(\boldsymbol{\rho}_i \circ \mathbf{d}_{k,n})\right\}. \quad (12)$$

These weights are used to construct the OW predictions for an initial date t_k and a lead time of h months:

$$[\mathbf{x}^{\mathrm{OW}}_{k,h}]_i = \sum_{n=1}^{N} w_{k,n,i}[\mathbf{x}^{f}_{k,h,n}]_i, \quad (13)$$

where $[\mathbf{x}]_i$ is the component of the vector $\mathbf{x}$ corresponding to the ith grid point.

To illustrate the outcome of this procedure, the optimal weights of four arbitrarily chosen members corresponding to one starting date are represented in Fig. A1 in the appendix.)

The localization radius L and the inflation factor λ° need to be set. Those variables are called "hyperparameters" and are tuned on the period 1987–2001. The sensitivity of the method to L and λ° is discussed in section 5a.

c. Validation metrics

For the sake of simplicity, hindcasts are denoted by $\mathbf{x}$ in this section, with no mention of the scheme used (OW, EW), the lead month, and the location. Two metrics are used to validate the hindcasts against the future observation $\mathbf{y}$, considered as ground truth. The metrics are the correlation C and the root-mean-square error (RMSE) defined as follows:

$$C = \frac{\sum(\mathbf{x} - \overline{\mathbf{x}})(\mathbf{y} - \overline{\mathbf{y}})}{\sqrt{\sum(\mathbf{x} - \overline{\mathbf{x}})^2\sum(\mathbf{y} - \overline{\mathbf{y}})^2}} \quad \text{and} \quad (14)$$

$$\mathrm{RMSE} = \sqrt{\overline{(\mathbf{x} - \mathbf{y})^2}}, \quad (15)$$

where the sum $\sum(\cdots)$ and the average $\overline{(\cdots)}$ can be performed either over a time period to provide maps, or over a spatial region to provide time series. If it is performed both spatially and temporally, it provides a global value. The spatial average is weighted to account for different cell areas. It is worth restating that here we analyze the SST anomaly (with the seasonal cycle removed).

The uncertainty of these scores is estimated by a bootstrapping method. For each score, an ensemble of 50 metrics is drawn by random sampling with replacement. Error bars are produced by considering the quantiles 0.1 and 0.9 of the ensemble. Differences are considered to be significant when they are of the same sign for more than 90% of the ensemble of metrics obtained by bootstrapping.

4. Results

a. Global analysis

In Fig. 1, the correlation between the observations and the hindcasts is computed globally as a function of lead time,

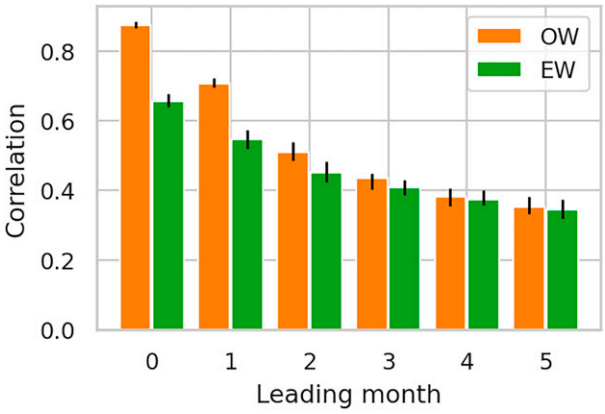

FIG. 1. Global correlation between the SST observations and the hindcast (1985–2010) with the optimal weights (in orange) and the equal weights (the baseline in green) method at different lead months. The error bars (black lines) defines the 10%–90% confidence interval estimated by bootstrapping.

which is defined in section 2b. It shows that correlation is significantly higher with the OW hindcasts than with the EW hindcasts up to a 2-month lead time. For a 0-month "lead" time, the OW uses future data (first week of the observation of the month following lead month 0), which explains the improvements compared to the analysis. It exemplifies nicely the benefit of the smoother method over filter approaches (Evensen and Van Leeuwen 2000). For a 1-month lead time, one week out of the four composing the monthly average has been used to determine the OW, so that the OW is not fully independent of the observations. It demonstrates that accounting for an extra week of observation with the weighting approach can improve significantly the forecast skill of the system.

If one sees the OW method as a "cheap" data assimilation procedure, the 1-month lead hindcast of the OW can be

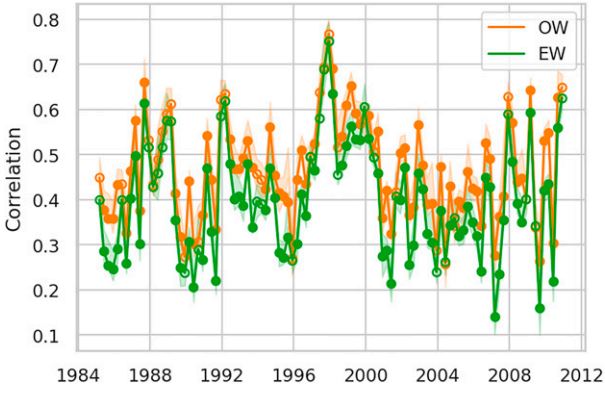

FIG. 2. Time evolution of the spatial correlation pattern between the SST observations and the hindcast computed with the optimal weights (in orange) and the equal weights (in green) method at 2-month lead time. The filled dots mean that the difference between OW and EW is significant, and the empty dots mean the difference is nonsignificant according to the criteria described in section 3c.

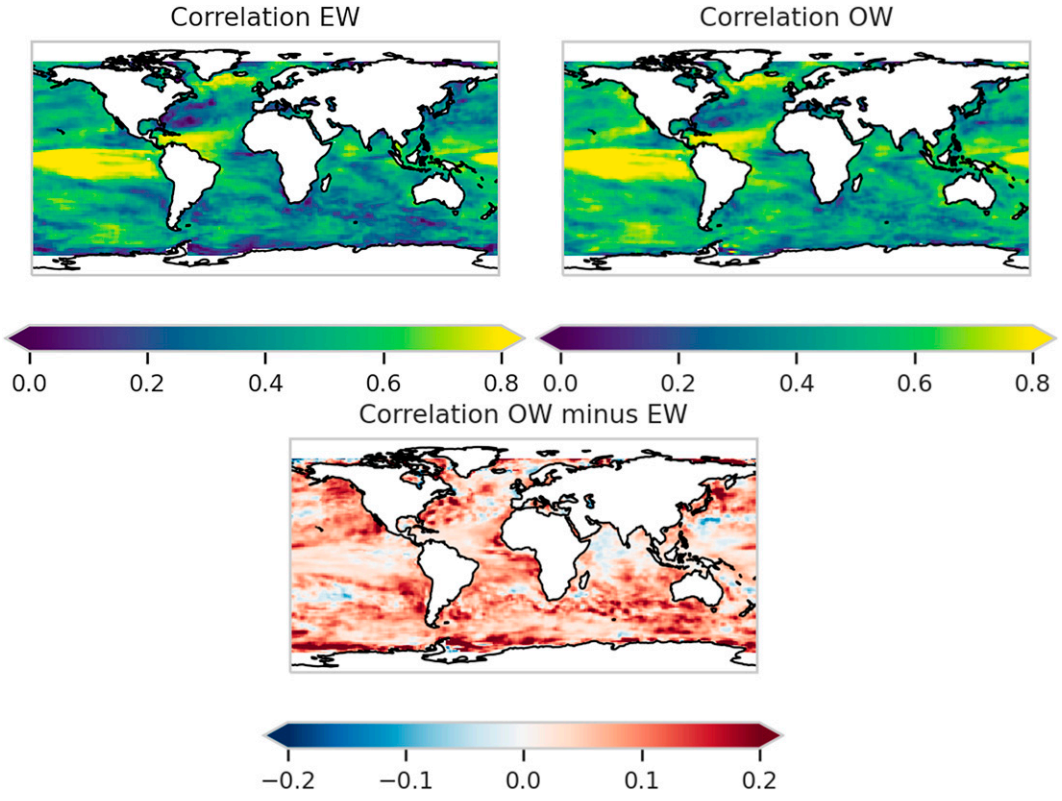

FIG. 3. Pointwise correlation at 2-month lead time between the SST observations and the hindcasts computed with (top left) equal weights and (top right) with OW. (bottom) The difference between the two, with positive values indicating that OW improves the correlation over the EW.

considered as an analysis with partial data availability (one week of data versus one month in the EnKF analysis). It can be seen in Fig. 1, that the correlation of the OW hindcast at a 1-month lead time is even slightly higher (0.71) than the EnKF analysis (0.66) at a 0-month lead time. The difference is small and most likely relates to the fact that OW is trained solely based on OISST, which is considered perfect for validation, while the EnKF analysis is based on several independent observations. The approach was tested with a simple Lorenz 1996 model (Lorenz and Emanuel 1998) and we found the OW at a lead time of 1 performs nearly as well as the analysis at a lead time of 0 (not shown). It is very encouraging that the OW can sustain a comparable level of accuracy as the analysis with just a week of additional data with a real prediction system as well.

Figure 2 shows that the improvements in the 2-month lead time spatial correlation pattern are consistent throughout the period of 1985–2010. No significant linear trends were found for both hindcasts. The OW hindcast performs significantly better than the EW hindcast over 71% of the time and is never significantly worse. Moreover, we can observe that the correlation is improved on average by 6.3×10^{-2} during the period used for tuning the hyperparameters (1987–2001), while it is 7.6×10^{-2} during the test period (2002–10). It suggests that the skill is sustained at a comparable level on a period independent of the tuning and gives confidence in the ability of our algorithm to generalize to future periods.

In Fig. 3, we show the pointwise correlation of the OW and EW hindcasts with SST observations at a 2-month lead time. The improvement of the OW over EW hindcasts is smaller in the regions where the EW already achieved good skill (equatorial Pacific, western tropical Atlantic, and the entrance of the Nordic seas). As for the global correlation skill, the OW algorithm is frequently significantly better (20% of the points) and is rarely worse (0.58% of the points) than the EW algorithm. We also assessed the skill of the OW et EW hindcast as a function of the starting date (January, April, July, or October): even if the correlation itself displays a seasonal variability, the incremental improvement provided by the OW algorithm is very stable throughout the year (not shown here) and does not depend significantly on the starting date. More generally, we would like to stress that the weighting procedure has the effect of improving an existing forecast system if newly available observations can be used; but of course, it does not prevent improving the forecasting system itself by other means (e.g., improving the model, assimilating other types of observations).

TABLE 1. Reliability metric for 2-month lead time hindcasts.

Hindcast	UMSE	Mean spread	r
EW	0.17	8.4×10^{-2}	8.4×10^{-2}
OW	0.15	6.3×10^{-2}	9.0×10^{-2}

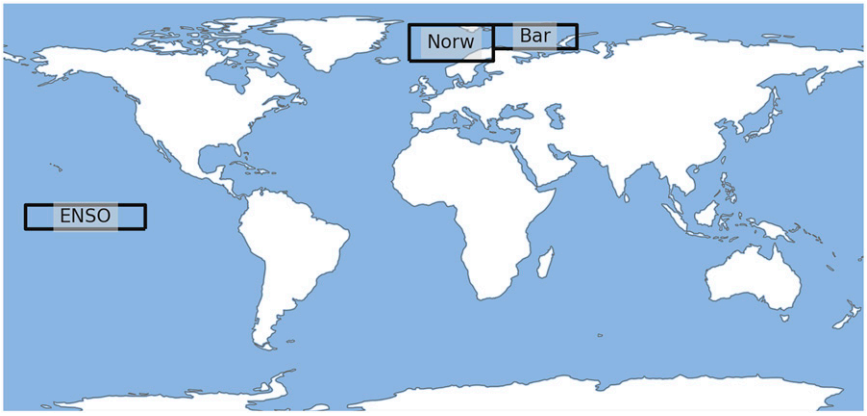

FIG. 4. Selected regions for the regional analysis. The abbreviations describing the regions are defined in Table 2.

b. *Reliability of the hindcast*

By modifying the weight of each member of an ensemble, the weighting procedure also modifies the spread s^2 of the ensemble defined as

$$s_m^2 = \frac{N+1}{N-1}\sum_{n=1}^{N} w_{m,n}(x_{m,n} - \overline{x}_m)^2, \tag{16}$$

where $\overline{x}_m = \sum_{n=1}^{N} w_n x_{m,n}$, $w_{m,n}$, and $x_{m,n}$ are the weight and the hindcast for the member n of a particular hindcast ensemble indexed by m, $1 < m < M$. The term M is the total number of considered hindcasts both in time and space. Note that the definition is valid for the OW hindcast with weights defined as in Eq. (13) as well as for the EW hindcast with constant weights $w_{m,n} = 1/N$. Following the formulation of Rodwell et al. (2016), we decompose the error in two terms:

$$\underbrace{\frac{1}{M-1}\sum_{m=1}^{M}(\overline{x}_m - y_m)^2 - \frac{1}{(M-1)M}\left(\sum_{m=1}^{M}(\overline{x}_m - y_m)\right)^2}_{\text{unbiased mean square error (UMSE)}}$$

$$= \underbrace{\frac{1}{M}\sum_{m=1}^{M} s_m^2}_{\text{mean spread}} + r, \tag{17}$$

where y_m is the observation corresponding to the hindcast $\overline{x}_m$ and r is a residual depending on the observation error and on the calibration error of the spread. The hindcast is said to be reliable if the residual is equal to the observation error, which means that it has a good estimate of its own uncertainty through the spread. It was shown in a previous work (Wang et al. 2017) that the NorCPM EW hindcast was mostly reliable but slightly underdispersive (i.e., overestimating r) in some regions. In Table 1, UMSE, the mean spread, and the residual r are shown for the EW and OW hindcast at a 2-month lead time. Values are averaged over all the grid points and all the dates. OW displays both a lower UMSE and a lower spread, which was expected since the OW hindcast was shown to be significantly better than the EW hindcast at a 2-month lead time. Nevertheless, the residual of the OW hindcast is 7.5% higher than the EW hindcast. It means that the OW hindcast has slightly degraded the reliability of the EW hindcast and that better skill of OW was achieved with a small underestimation of the spread. It is not surprising since the weighting procedure discards members that are far from the observation so that the number of effective members is lower and enhances sampling error. As for the EnKF data assimilation (e.g., Anderson 2001; Raanes et al. 2019), sampling error causes a spurious reduction of the ensemble spread and a degradation of the reliability. Nevertheless, we highlight that the degradation is very small and that the ensemble spread of the OW hindcast can still be used to assess uncertainty.

c. *Regional analysis*

We also assess the impact in specific regions. We have selected three regions highlighted in Fig. 4: The Norwegian Sea (Norw), the Barents Sea (Bar), and the ENSO region. In Table 2, the correlation and RMSE value of the 2-month lead time hindcasts are reported globally and in the selected regions. For the global ocean and ENSO, the OW hindcast significantly outperforms the EW for correlation and RMSE following the bootstrapping criteria described in section 3c. For the Barents Sea, OW gives a higher correlation and a lower RMSE for more than 80% of the bootstrap samples but slightly below the significance threshold set to 90%. In the Norwegian Sea, there are no significant differences between the EW and the OW hindcasts in terms of correlation and a nonsignificant degradation in terms of RMSE.

TABLE 2. Correlation and RMSE at a 2-month lead time for selected regions (bold font highlights the best score).

		Correlation		RMSE	
Region	Abbreviation	EW	OW	EW	OW
Global		0.45	**0.51**	0.54	**0.52**
ENSO-3.4 index	ENSO	0.90	**0.93**	0.44	**0.36**
Barents Sea	Bar	0.38	**0.48**	0.61	**0.58**
Norwegian Sea	Norw	0.63	0.64	**0.38**	0.43

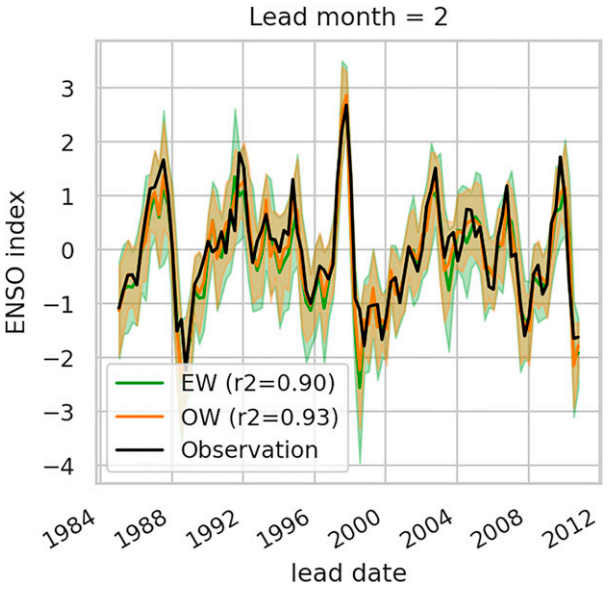

FIG. 5. ENSO-3.4 index for the EW hindcast (in green), the OW hindcast (in orange), and the observations (in black). The shaded areas represent the hindcast ± 2 times the spread.

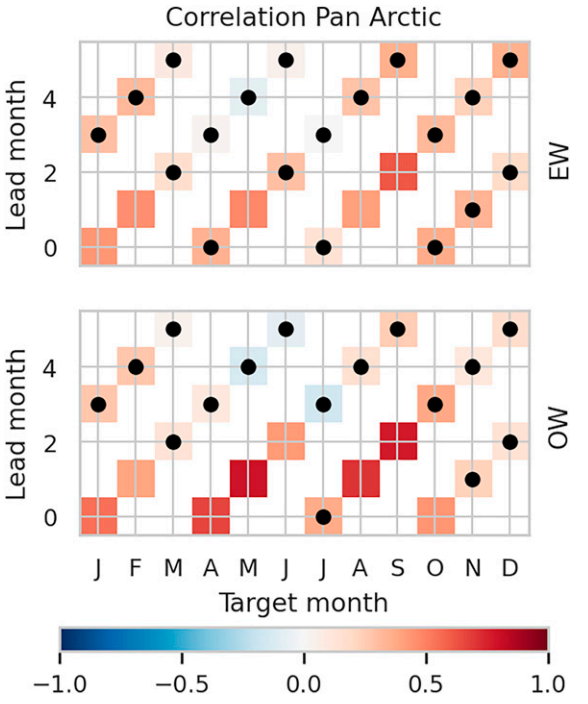

FIG. 6. Correlation coefficients between detrended sea ice extent from the observations in the Arctic and the hindcast from (top) EW and (bottom) OW. Dots correspond to correlations that are not statistically significant at a 95% confidence level.

To illustrate the regional differences between the EW and the OW hindcast, we present in Fig. 5 the time series of the Niño-3.4 index (SST anomaly in the region, 5°S–5°N and 120°–170°W) for the EW hindcast, the OW hindcast and the observations at a 2-month lead time. Both the EW and OW hindcasts predict well the observations which lay well within 2 standard deviations. Most of the time the OW and EW means overlap each other and are barely discernible. Nevertheless, we can see that the OW captures slightly better the peaks (El Niño and La Niña) that are sometime too early, too low, or too strong in EW. Overall the OW prediction is closer to the observation for 63% of the hindcasts. We can also see that the reliability of the ensemble is well preserved and that the spread reduction observed in section 4b is very moderate.

d. *Sea ice extent*

Here, we assess the benefit of the optimal weights estimated using the new SST observations on another variable, e.g., sea ice. Sea ice concentration and SST are anticorrelated (Wang et al. 2019), so we expect that the improved skill for SST will also yield improved skill for the sea ice concentration. Furthermore, as the ocean data are used to update the sea ice component (Bethke et al. 2021), we expect the dynamical consistency between the ocean and sea ice component to be in good agreement with the forecasted ensemble. In Fig. 6, we show the skill of predicting the sea ice extent in the Arctic between the EW and OW forecast. Sea ice extent is defined as the area of grid cells where the sea ice concentration is greater than 15%. As we are primarily interested in interannual variability, we have analyzed the detrended sea ice extent, as in Wang et al. (2019), Bushuk et al. (2017), Kimmritz et al. (2019). The sea ice extent hindcast is validated against the one computed from sea ice concentration observations provided by HadISST2.1.0.0 (Rayner et al. 2003). It can be seen that, even if sea ice concentration observations were not used to determine the weights, the OW procedure leads to a better skill of the sea ice extent in particular during the transition months: freezing (around September) and melting (around May). During these transition periods, changes are relatively quick as the influence of atmospheric variability becomes more predominant (Bushuk et al. 2017; Dai et al. 2020; Stroeve et al. 2014; Kwok and Rothrock 2009; Maslanik et al. 2011). Having access to one week of fresh data can therefore be determinant. For the starting month in July, the OW correlation outperforms EW for lead times of 2 and 3 months in September. Predicting the decline in September sea ice extent is important for the ecosystem, local communities, and economic activities in the Arctic, such as tourism, fisheries, shipping, and resource exploitation (e.g., Liu and Kronbak 2010). Similarly, for a starting month in April, the skill of OW is significant up to a 2-month lead time while the skill is only significant (and smaller) for a 1-month lead time in the EW scheme.

This demonstrates the potential of the method to predict variables that were not used to constrain the weights and provide consistent hindcasts across several variables and components.

5. Discussion

a. *Tuning hyperparameters*

The OW method relies on two hyperparameters: the localization radius L and the inflation factor λ°. These parameters

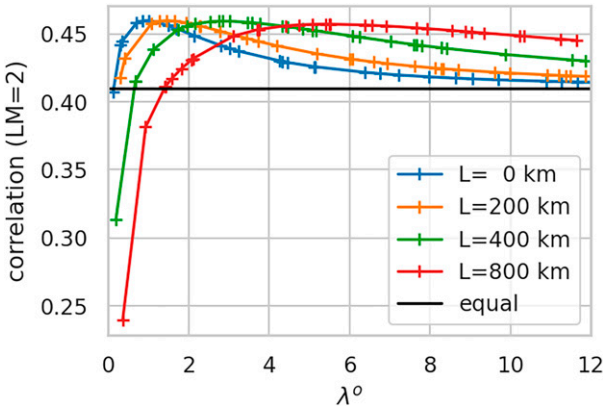

FIG. 7. Global correlation between SST observations and the optimal weight hindcasts at a 2-month lead for different values of inflation λ° and localization radius ($L = 0$ km in blue, $L = 200$ km in orange, $L = 400$ km in green, and $L = 800$ km in red). The "+" markers indicate the values for which the correlation was computed. The black line reports the performance of the EW hindcasts. The correlation is computed globally on the whole ocean and for hindcast starting from January 1987 to October 2000.

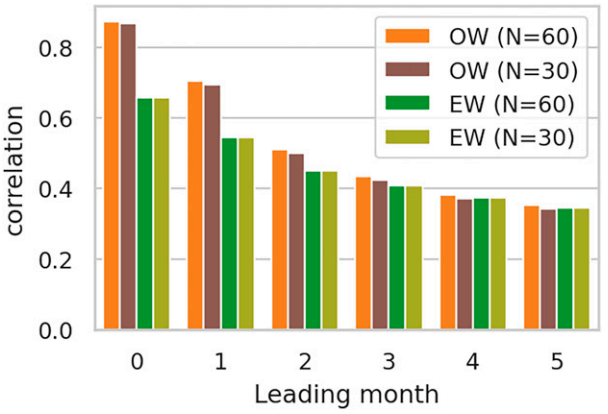

FIG. 8. Global correlation between the SST observations and the OW hindcasts when computed with the full ensemble of $N = 60$ members (in orange) and with a reduced ensemble of $N = 30$ members (in brown), as a function of lead time. The EW hindcast correlation is represented in dark green when computed with $N = 60$ and represented in light green when computed with $N = 30$. The correlation is computed globally for hindcast started from January 1985 to October 2010.

have been tuned within acceptable ranges. Four different values of the localization radius have been tested: L (km) $\in$ {0, 200, 400, 800}, which are typical ranges used in ocean data assimilation systems (Sakov et al. 2012; Massonnet et al. 2014; Schiller et al. 2020; Lellouche et al. 2013). For each localization radius value, we explored 30 inflation factors between 0.1 and 14. The optimal value has been selected based on global correlation at a 2-month lead time. The optimization was done using the Hyperopt Python package (Bergstra et al. 2013) that implements Bayesian optimization using the tree-structured Parzen estimator approach (Bergstra et al. 2011). The principle of the algorithm is to iteratively update a probability distribution of the hindcast performance over the hyperparameters, and use this distribution to guide the selection of the next set of hyperparameters to evaluate. Hence, we can efficiently search the hyperparameter for values that are more likely to improve performance while controlling the number of different values to test ("+" markers in Fig. 7).

In Fig. 7, the global correlation at a 2-month lead time is computed as a function of the localization radius and the inflation. Except for very small inflation ($\lambda^\circ < 1$), the correlation of the OW hindcast is better than that of the EW. It should be emphasized that $\lambda^\circ < 1$ counteracts the purpose of inflation (deflation) to prevent filter degeneracy. It is also noted that for each explored value of the localization radius, there are optimal values of λ° which achieve a correlation comparable to the optimal setting. This includes the hyper-localization version ($L = 0$). It implies that we could have thus tuned only one hyperparameter and that matching quantitative accuracy can be achieved without localization. A larger localization only ensures smoothness in the output. This is clearly an advantage as it avoids introducing dynamical imbalance in the prediction. A larger localization radius is also associated with a larger optimal inflation factor. This is expected since the dimension of the innovation defined in Eq. (10) increases with the localization radius. As such, the particle filter is more subject to degeneracy, which must be mitigated by the inflation factor. Another interesting point is that for a larger localization radius, the sensitivity to the inflation factors is reduced (optimal range of λ° is broadening). The final parameters presented in this work were $L = 400$ km and $\lambda^\circ = 2.84$.

We have designed the optimal hyperparameters to be constant in space and time (which, of course, does not mean that the weights are constant). In classical ocean assimilation systems, localization is fixed in time but can vary spatially—e.g., latitude dependent (Zhang et al. 2005; Wang et al. 2017). Due to practical constraints related to development time, it was not implemented in this first version of the OW algorithm, but it would be interesting to test whether the performance would improve with a spatially varying localization that matches that of the assimilation system. The inflation is most often constant in space and time (e.g., Sakov et al. 2012), even if adaptive schemes are gaining in popularity (Anderson 2009; El Gharamti et al. 2019). One could extend the degree of freedom of the hyperparameters space to allow a search in space and time, but it could lead to overfitting and impair the generalization skill of our model. Given the limited number of data used to tune the hyperparameters, it is safe to limit the dimension of the search space for those parameters to prioritize the generalization skill over the accuracy.

b. *Ensemble size*

Another key parameter of the method is the size of the ensemble. The computational cost of producing the dynamical forecast (which dominates the overall cost) increases linearly

with the ensemble size. We benefit from a 60-member ensemble, but it could be worth applying our algorithm to a smaller ensemble. We performed a sensitivity experiment by randomly removing 30 members from the ensemble and recomputing the OW with a reduced ensemble. In Fig. 8, we present the global correlation with the OW and EW method as a function of the lead time for the original ensemble size of 60 members and for an ensemble of size 30. It appears that the skill is marginally reduced with half the ensemble size, and it is still substantially better than the EW performance, also reported in Fig. 8, computed with both 30 and 60 members. It shows that the algorithm is still applicable to a smaller ensemble.

It should be emphasized that the sensitivity to ensemble size presented here will differ if we test the approach on a different metric. Atmospheric quantities require a larger ensemble size because it is more sensitive to initial conditions on these time scales, and an ensemble size larger than 40 members are needed to predict skillfully winter North Atlantic Oscillation (Dunstone et al. 2016).

The OW enables the possibility to incorporate new observations by computing weights following Eq. (12), which is cheaper than recomputing the ensemble forecast, so some computing cost is saved. Therefore, instead of allocating computing resources to compute more frequent ensemble forecasts (but with a smaller ensemble), it is possible to compute bigger ensembles.

6. Conclusions and perspectives

We have presented an algorithm that can improve the accuracy of an ensemble forecast by utilizing newly available observations that were not used for producing the prediction. The observations are used to estimate weights for the ensemble forecast. The algorithm is computationally cheap, time efficient (run on a laptop in a few minutes), and easy to implement. In the case of the Norwegian Climate Prediction Model, a week of SST observations is available at the time of the forecast delivery in an operational setting, and we have shown that it improves the prediction accuracy significantly up to a 2-month lead time, with a global mean correlation of 0.51 against 0.45 for the equal weight predictions. The improvement is significant regionally, such as in the ENSO region, or to a lesser extent in the Barents Sea. Our algorithm achieves optimal performance by tuning only one hyperparameter and is likely to be able to generalize to future data. Weights determined using SST data can also be used to improve the skill of other quantities, such as sea ice extent.

The results presented here demonstrate the potential of the method to enhance the accuracy of our operational forecast, and sustain a high level of accuracy in between the production cycle. Still, the algorithm presented could be further refined. For example, we could easily adapt the algorithm to consider all daily observations unused instead of the weekly average. These modifications could even improve the results since the model outputs and the observations would be more in phase and remove the approximation about time resolution currently made.

The algorithm presented here has been demonstrated using sea surface temperature observations to determine the weights. Nevertheless, the approach can be generalized to other observations or even by considering short-range model forecasts (provided by numerical weather predictions) as observations. It can also provide a continuously up-to-date forecast in between the dynamical production steps, at the time needed by the stakeholder.

Finally, we have used an ensemble issued by one model, but future development could be to leverage a multimodel ensemble to improve the seasonal forecast skill.

Acknowledgments. This work was supported by the Norwegian Research Council Projects (270733, 328886, and 309562) and by the Trond Mohn Foundation, under Project BFS2018TMT01. This work has also received a grant for computer time from the Norwegian Program for supercomputing (NOTUR2, Project nn9039k) and a storage grant (NORSTORE, NS9039k). The authors warmly thank Stephen Outten (NERSC) for his contribution to improving the manuscript. Julien Brajard is also an associate professor at Sorbonne Université (Paris, France).

Data availability statement. The NOAA OISSTV2 data used for validation are available at ftp://ftp.cdc.noaa.gov/Datasets/noaa.oisst.v2/. Hindcast SST data used in this article are available at https://archive.sigma2.no/pages/public/datasetDetail.jsf?id=10.11582/2023.00045 and are associated with https://doi.org/10.11582/2023.00045. The data are compressed into a unique archive file (12.42 GB).

APPENDIX

Example of Optimal Weights

To illustrate how are the weights computed by our algorithm, we have represented a set of weight values in Fig. A1 for a given starting date of January 1987 and for 4 arbitrarily chosen members (over 60). As a comparison, each EW hindcast's weight is equal to 1/60.

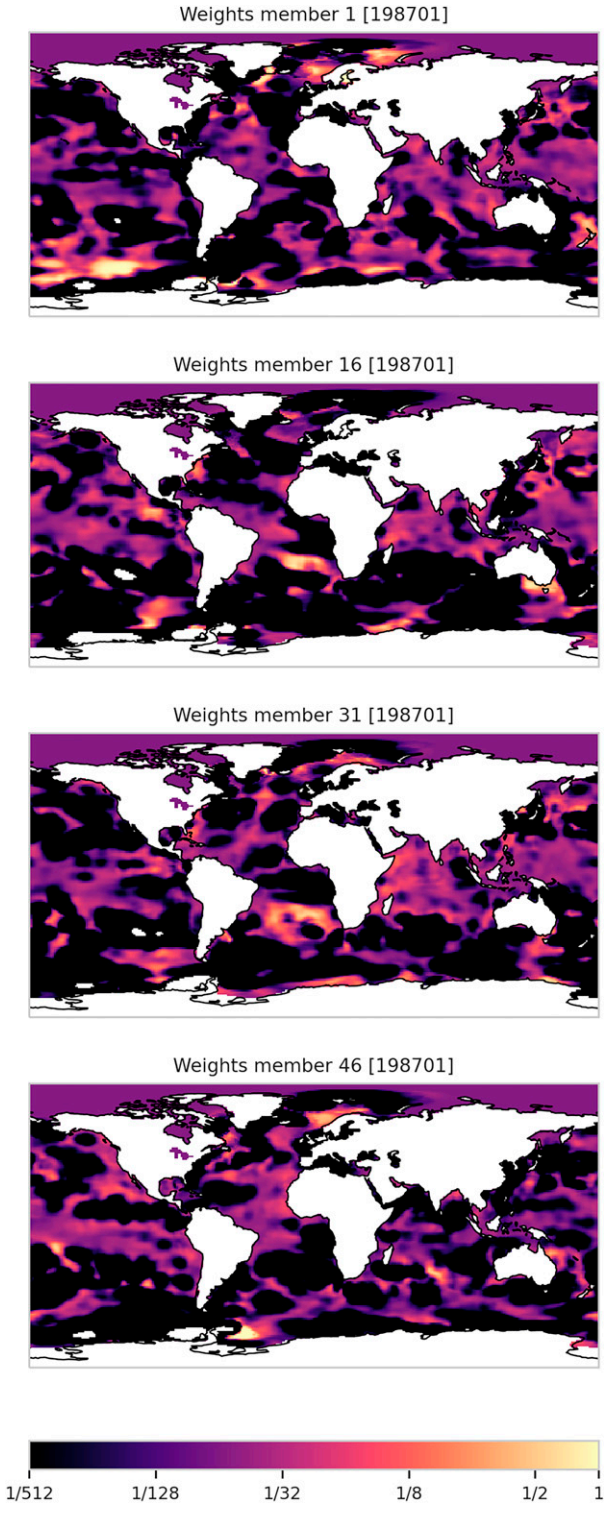

FIG. A1. Example of weights for four arbitrary members (number 1, 16, 31, 46) determined for the starting date January 1987. The weights have been determined using the first week of February 1987 as observations. To enhance the contrast, the color bar is on a logarithmic scale.

REFERENCES

Anderson, J. L., 2001: An ensemble adjustment Kalman filter for data assimilation. *Mon. Wea. Rev.*, **129**, 2884–2903, https://doi.org/10.1175/1520-0493(2001)129<2884:AEAKFF>2.0.CO;2.

——, 2009: Spatially and temporally varying adaptive covariance inflation for ensemble filters. *Tellus*, **61A**, 72–83, https://doi.org/10.1111/j.1600-0870.2008.00361.x.

Becker, E. J., B. P. Kirtman, M. L'Heureux, Á. G. Muñoz, and K. Pegion, 2022: A decade of the North American Multimodel Ensemble (NMME): Research, application, and future directions. *Bull. Amer. Meteor. Soc.*, **103**, E973–E995, https://doi.org/10.1175/BAMS-D-20-0327.1.

Bentsen, M., and Coauthors, 2013: The Norwegian Earth System Model, NorESM1-M—Part 1: Description and basic evaluation of the physical climate. *Geosci. Model Dev.*, **6**, 687–720, https://doi.org/10.5194/gmd-6-687-2013.

Bergstra, J., R. Bardenet, Y. Bengio, and B. Kégl, 2011: Algorithms for hyper-parameter optimization. *NIPS'11: Proc. 24th Int. Conf. on Neural Information Processing Systems*, Granada, Spain, Association for Computing Machinery, 2546–2554, https://dl.acm.org/doi/10.5555/2986459.2986743.

——, D. Yamins, and D. Cox, 2013: Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. *Proc. 30th Int. Conf. Machine Learning*, Atlanta, GA, PMLR, 115–123, https://proceedings.mlr.press/v28/bergstra13.html.

Bethke, I., and Coauthors, 2021: NorCPM1 and its contribution to CMIP6 DCPP. *Geosci. Model Dev.*, **14**, 7073–7116, https://doi.org/10.5194/gmd-14-7073-2021.

Bushuk, M., R. Msadek, M. Winton, G. A. Vecchi, R. Gudgel, A. Rosati, and X. Yang, 2017: Skillful regional prediction of Arctic sea ice on seasonal timescales. *Geophys. Res. Lett.*, **44**, 4953–4964, https://doi.org/10.1002/2017GL073155.

Carrassi, A., M. Bocquet, L. Bertino, and G. Evensen, 2018: Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *Wiley Interdiscip. Rev.: Climate Change*, **9**, e535, https://doi.org/10.1002/wcc.535.

Dai, P., Y. Gao, F. Counillon, Y. Wang, M. Kimmritz, and H. R. Langehaug, 2020: Seasonal to decadal predictions of regional Arctic sea ice by assimilating sea surface temperature in the Norwegian climate prediction model. *Climate Dyn.*, **54**, 3863–3878, https://doi.org/10.1007/s00382-020-05196-4.

Dobrynin, M., and Coauthors, 2018: Improved teleconnection-based dynamical seasonal predictions of boreal winter. *Geophys. Res. Lett.*, **45**, 3605–3614, https://doi.org/10.1002/2018GL077209.

Dunstone, N., D. Smith, A. Scaife, L. Hermanson, R. Eade, N. Robinson, M. Andrews, and J. Knight, 2016: Skilful predictions of the winter North Atlantic Oscillation one year ahead. *Nat. Geosci.*, **9**, 809–814, https://doi.org/10.1038/ngeo2824.

El Gharamti, M., K. Raeder, J. Anderson, and X. Wang, 2019: Comparing adaptive prior and posterior inflation for ensemble filters using an atmospheric general circulation model. *Mon. Wea. Rev.*, **147**, 2535–2553, https://doi.org/10.1175/MWR-D-18-0389.1.

Evensen, G., 2003: The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dyn.*, **53**, 343–367, https://doi.org/10.1007/s10236-003-0036-9.

——, and P. J. van Leeuwen, 2000: An ensemble Kalman smoother for nonlinear dynamics. *Mon. Wea. Rev.*, **128**, 1852–1867, https://doi.org/10.1175/1520-0493(2000)128<1852:AEKSFN>2.0.CO;2.

——, F. C. Vossepoel, and P. J. van Leeuwen, 2022: *Data Assimilation Fundamentals: A Unified Formulation of the State and Parameter Estimation Problem*. Springer, 246 pp.

Gaspari, G., and S. E. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Quart. J. Roy. Meteor. Soc.*, **125**, 723–757, https://doi.org/10.1002/qj.49712555417.

Good, S. A., M. J. Martin, and N. A. Rayner, 2013: EN4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. *J. Geophys. Res. Oceans*, **118**, 6704–6716, https://doi.org/10.1002/2013JC009067.

Goutham, N., R. Plougonven, H. Omrani, S. Parey, P. Tankov, A. Tantet, P. Hitchcock, and P. Drobinski, 2022: How skillful are the European subseasonal predictions of wind speed and surface temperature? *Mon. Wea. Rev.*, **150**, 1621–1637, https://doi.org/10.1175/MWR-D-21-0207.1.

Hewitt, C. D., and J. A. Lowe, 2018: Toward a European climate prediction system. *Bull. Amer. Meteor. Soc.*, **99**, 1997–2001, https://doi.org/10.1175/BAMS-D-18-0022.1.

Kimmritz, M., F. Counillon, L. H. Smedsrud, I. Bethke, N. Keenlyside, F. Ogawa, and Y. Wang, 2019: Impact of ocean and sea ice initialisation on seasonal prediction skill in the Arctic. *J. Adv. Model. Earth Syst.*, **11**, 4147–4166, https://doi.org/10.1029/2019MS001825.

Kirtman, B. P., and Coauthors, 2014: The North American multimodel ensemble: Phase-1 seasonal-to-interannual prediction; Phase-2 toward developing intraseasonal prediction. *Bull. Amer. Meteor. Soc.*, **95**, 585–601, https://doi.org/10.1175/BAMS-D-12-00050.1.

Kwok, R., and D. Rothrock, 2009: Decline in Arctic sea ice thickness from submarine and ICESat records: 1958–2008. *Geophys. Res. Lett.*, **36**, L15501, https://doi.org/10.1029/2009GL039035.

Lean, P., E. V. Hólm, M. Bonavita, N. Bormann, A. P. McNally, and H. Järvinen, 2021: Continuous data assimilation for global numerical weather prediction. *Quart. J. Roy. Meteor. Soc.*, **147**, 273–288, https://doi.org/10.1002/qj.3917.

Lellouche, J.-M., and Coauthors, 2013: Evaluation of global monitoring and forecasting systems at Mercator Océan. *Ocean Sci.*, **9**, 57–81, https://doi.org/10.5194/os-9-57-2013.

Liu, M., and J. Kronbak, 2010: The potential economic viability of using the Northern Sea Route (NSR) as an alternative route between Asia and Europe. *J. Transp. Geogr.*, **18**, 434–444, https://doi.org/10.1016/j.jtrangeo.2009.08.004.

Lorenz, E. N., and K. A. Emanuel, 1998: Optimal sites for supplementary weather observations: Simulation with a small model. *J. Atmos. Sci.*, **55**, 399–414, https://doi.org/10.1175/1520-0469(1998)055<0399:OSFSWO>2.0.CO;2.

Mariotti, A., and Coauthors, 2020: Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond. *Bull. Amer. Meteor. Soc.*, **101**, E608–E625, https://doi.org/10.1175/BAMS-D-18-0326.1.

Maslanik, J., J. Stroeve, C. Fowler, and W. Emery, 2011: Distribution and trends in Arctic sea ice age through spring 2011. *Geophys. Res. Lett.*, **38**, L13502, https://doi.org/10.1029/2011GL047735.

Massonnet, F., H. Goosse, T. Fichefet, and F. Counillon, 2014: Calibration of sea ice dynamic parameters in an ocean-sea ice model using an ensemble Kalman filter. *J. Geophys. Res. Oceans*, **119**, 4168–4184, https://doi.org/10.1002/2013JC009705.

Meehl, G. A., and Coauthors, 2021: Initialized earth system prediction from subseasonal to decadal timescales. *Nat. Rev. Earth Environ.*, **2**, 340–357, https://doi.org/10.1038/s43017-021-00155-x.

Palmer, T., A. Döring, and G. Seregin, 2014: The real butterfly effect. *Nonlinearity*, **27**, R123–R141, https://doi.org/10.1088/0951-7715/27/9/R123.

Raanes, P. N., M. Bocquet, and A. Carrassi, 2019: Adaptive covariance inflation in the ensemble Kalman filter by Gaussian scale mixtures. *Quart. J. Roy. Meteor. Soc.*, **145**, 53–75, https://doi.org/10.1002/qj.3386.

Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan, 2003: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.*, **108**, 4407, https://doi.org/10.1029/2002JD002670.

Reynolds, R. W., N. A. Rayner, T. M. Smith, D. C. Stokes, and W. Wang, 2002: An improved in situ and satellite SST analysis for climate. *J. Climate*, **15**, 1609–1625, https://doi.org/10.1175/1520-0442(2002)015<1609:AIISAS>2.0.CO;2.

Rodwell, M. J., S. Lang, N. B. Ingleby, N. Bormann, E. Hólm, F. Rabier, D. Richardson, and M. Yamaguchi, 2016: Reliability in ensemble data assimilation. *Quart. J. Roy. Meteor. Soc.*, **142**, 443–454, https://doi.org/10.1002/qj.2663.

Sakov, P., F. Counillon, L. Bertino, K. A. Lisæter, P. R. Oke, and A. Korablev, 2012: TOPAZ4: An ocean-sea ice data assimilation system for the North Atlantic and Arctic. *Ocean Sci.*, **8**, 633–656, https://doi.org/10.5194/os-8-633-2012.

Schiller, A., and Coauthors, 2020: Bluelink Ocean forecasting Australia: 15 years of operational ocean service delivery with societal, economic and environmental benefits. *J. Oper. Oceanogr.*, **13** (1), 1–18, https://doi.org/10.1080/1755876X.2019.1685834.

Shukla, J., 1998: Predictability in the midst of chaos: A scientific basis for climate forecasting. *Science*, **282**, 728–731, https://doi.org/10.1126/science.282.5389.728.

Snyder, C., T. Bengtsson, P. Bickel, and J. Anderson, 2008: Obstacles to high-dimensional particle filtering. *Mon. Wea. Rev.*, **136**, 4629–4640, https://doi.org/10.1175/2008MWR2529.1.

Stroeve, J. C., T. Markus, L. Boisvert, J. Miller, and A. Barrett, 2014: Changes in Arctic melt season and implications for sea ice loss. *Geophys. Res. Lett.*, **41**, 1216–1225, https://doi.org/10.1002/2013GL058951.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, **93**, 485–498, https://doi.org/10.1175/BAMS-D-11-00094.1.

Thorey, J., V. Mallet, and P. Baudin, 2017: Online learning with the continuous ranked probability score for ensemble forecasting. *Quart. J. Roy. Meteor. Soc.*, **143**, 521–529, https://doi.org/10.1002/qj.2940.

van Leeuwen, P. J., H. R. Künsch, L. Nerger, R. Potthast, and S. Reich, 2019: Particle filters for high-dimensional geoscience applications: A review. *Quart. J. Roy. Meteor. Soc.*, **145**, 2335–2365, https://doi.org/10.1002/qj.3551.

Vitart, F., and A. W. Robertson, 2018: The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events. *npj Climate Atmos. Sci.*, **1**, 3, https://doi.org/10.1038/s41612-018-0013-0.

——, and Coauthors, 2017: The subseasonal to seasonal (S2S) prediction project database. *Bull. Amer. Meteor. Soc.*, **98**, 163–173, https://doi.org/10.1175/BAMS-D-16-0017.1.

Wang, Y., F. Counillon, I. Bethke, N. Keenlyside, M. Bocquet, and M. Shen, 2017: Optimising assimilation of hydrographic profiles into isopycnal ocean models with ensemble data assimilation. *Ocean Modell.*, **114**, 33–44, https://doi.org/10.1016/j.ocemod.2017.04.007.

——, ——, N. Keenlyside, L. Svendsen, S. Gleixner, M. Kimmritz, P. Dai, and Y. Gao, 2019: Seasonal predictions initialised by

assimilating sea surface temperature observations with the EnKF. *Climate Dyn.*, **53**, 5777–5797, https://doi.org/10.1007/s00382-019-04897-9.

Zhang, F., Y. Q. Sun, L. Magnusson, R. Buizza, S.-J. Lin, J.-H. Chen, and K. Emanuel, 2019: What is the predictability limit of midlatitude weather? *J. Atmos. Sci.*, **76**, 1077–1091, https://doi.org/10.1175/JAS-D-18-0269.1.

Zhang, S., M. J. Harrison, A. T. Wittenberg, A. Rosati, J. L. Anderson, and V. Balaji, 2005: Initialization of an ENSO forecast system using a parallelized ensemble filter. *Mon. Wea. Rev.*, **133**, 3176–3201, https://doi.org/10.1175/MWR3024.1.

Zhu, J., A. Kumar, H.-C. Lee, and H. Wang, 2017: Seasonal predictions using a simple ocean initialization scheme. *Climate Dyn.*, **49**, 3989–4007, https://doi.org/10.1007/s00382-017-3556-6.