https://doi.org/10.5194/egusphere-2023-2649
Preprint. Discussion started: 1 December 2023

Selecting and weighting dynamical models using data-driven approaches

Pierre Le Bras[1,2,3], Florian Sévellec[1,3], Pierre Tandeo[2,3], Juan Ruiz[4,5,6], and Pierre Ailliot[7]

[1]Laboratoire d'Océanographie Physique et Spatiale, IUEM, Univ Brest CNRS IRD Ifremer, Brest, France
[2]IMT Atlantique, Lab-STICC, UMR CNRS 6285, F-29238, France
[3]Odyssey team-project, INRIA IMT Atlantique CNRS, France
[4]Facultad de Ciencias Exactas y Naturales, Departamento de Ciencias de la Atmósfera y los Océanos, Universidad de Buenos Aires, Buenos Aires, Argentina
[5]Centro de Investigaciones del Mar y la Atmósfera (CIMA), CONICET–Universidad de Buenos Aires, Buenos Aires, Argentina
[6]Instituto Franco-Argentino para el Estudio del Clima y sus Impactos (IRL IFAECI), CNRS–IRD–CONICET–UBA, Buenos Aires, Argentina
[7]Univ Brest, CNRS UMR 6205, Laboratoire de Mathematiques de Bretagne Atlantique, France

Correspondence: Pierre Le Bras (pierre.lebras@univ-brest.fr)

Abstract. In geosciences, multi-model ensembles are helpful to explore the robustness of a range of results. To obtain a synthetic and improved representation of the studied dynamic system, the models are usually weighted. The simplest method, namely the model democracy, gives equal weights to all models, while more advanced approaches base weights on agreement with available observations. Here, we focus on determining weights for various versions of an idealized model of Atlantic Meridional Overturning Circulation. This is done by assessing their performance against synthetic observations (generated from one of the model versions) within a data assimilation framework using EnKF. In contrast to traditional data assimilation, we implement data-driven forecasts using the analog method based on catalogs of short-term trajectories. This approach allows us to efficiently emulate the model's dynamics while keeping computational costs low. For each model version, we compute a local performance metric, known as the contextual model evidence, to compare observations and model forecasts. This metric, based on the innovation likelihood, is sensitive to differences in model dynamics and considers forecast and observation uncertainties. Finally, the weights are calculated using both model performance and model codependency, and then evaluated on climatologies of long-term simulations. Results show good performance in identifying numerical simulations that best replicate observed short-term variations. Additionally, it outperforms benchmark approaches such as model democracy or climatologies-based strategies when reconstructing missing distributions. These findings encourage the application of the proposed methodology to more complex datasets in the future, like climate simulations.

1 Introduction

In the geosciences, several numerical models are usually available to represent the same complex system. For example, the Earth's global climate system is implemented in a range of different numerical models that are, for some, gathered within the Couple Model Intercomparison Projet (CMIP, Eyring et al. 2016). Due to the different parameterizations (related to unresolved processes and numerical implementations) and given particular structural equations, a model represents an imperfect image of the studied system. In other words, each model is characterized by its own strengths and weaknesses, which impact its predictive skills. The use of the multi-model ensemble (MME) to predict a specific climate variable is then considered more informative than any individual model. Here, the models are assumed to complement each other using MME which has the potential to improve climate system representation (Abramowitz et al. 2019).

Nevertheless it remains the question of how to gather the information produces by the range of models. In this context it is usual to follow a model democracy approach, giving equal weight to all models to enhance the representativeness. This is what is mostly done in the 6th Assessment Report (AR6) of the Intergovernmental Panel on Climate Change (IPCC 2021). However, this strategy has been debated in the community due to the strong assumptions it implies: equal probability of all models to represent all variables in the whole phase-space of the system and independence between them (e.g., Knutti 2010; Sanderson et al. 2015; Knutti et al. 2019).

A trade-off between MME-based model democracy and selecting a single model would be to weight-average the models. The weights can be found through model performance, which consists in evaluating the consistency between model simulations and available observations. This procedure should be adapted to the specificities of model simulations being processed (Eyring et al. 2019). Hence, the dynamic behaviour of the system needs to be properly taken into account. This can be done, for example, by assessing the models' ability to describe successive sequences of observations. In addition, the weights should account for the degree of dependency between the models, as they usually share some elements in their structure or parameterizations (Knutti et al. 2010; Abramowitz et al. 2019). Several approaches have been suggested to measure the model-observations consistency and thus obtain the individual model weights.

The first approaches are based on descriptive statistics. This is the case of "reliability ensemble averaging" (Giorgi and Mearns 2002) and the "ClimWIP method" (Climate model Weighting by Independence and Performance, Sanderson et al. 2015 and Knutti et al. 2017). In those two methods, weights are based on the ability of model simulations to reproduce observed statistical diagnostics for one or more variables (e.g., mean, variability, or trend fields), while including a component that measures inter-model dependence. However, the score formalism does not evaluate how the models reproduce sequences of observations (which reflect model dynamics). ClimWIP method was applied in numerous studies at global or regional spatial-scales with a range of climate variables such as atmospheric temperatures or precipitations (Sanderson et al. 2015; Sanderson et al. 2017; Brunner et al. 2019; Brunner et al. 2020; Merrifield et al. 2020), arctic sea-ice (Knutti et al. 2017), or antarctic ozone depletion (Amos et al. 2020).

Second approaches are included in a probabilistic Bayesian framework, such as "Bayesian model averaging" (BMA, e.g., Raftery et al. 2005; Min et al. 2007; Sexton et al. 2012; Olson et al. 2016; Olson et al. 2019) and "kriging methods" applied at global and local scale (Ribes et al. 2021 and Ribes et al. 2022, respectively). In BMA, the distributions of model outputs are updated using available observations to obtain a posterior weighted distribution. The weights for candidate models are determined using Bayes' theorem which combines prior beliefs (typically uniform weights) with the global likelihood of the observations estimated for each model. These weights reflect the confidence in each model considering both the data and uncertainty. However, evaluating model consistency over time is tricky because the distributions involve time aggregated quantities on large time periods, making it less suitable for exploiting information about model dynamics.

In this study, we evaluate the performance of competing models based on their ability to reproduce short-term dynamics of the system described by observed sequences. To achieve this, a data-driven approach is adopted within a data assimilation framework (Ruiz et al. 2022). One of the main advantage is the use of already existing simulations to produce free-model probabilistic forecasts (i.e. without having to rerun the models, which greatly reduces the computational cost), while retaining the useful properties of data assimilation (Lguensat et al. 2017). The data-driven forecasts are estimated using regression based on the classic least squares, but applied locally in the phase space, to capture the nonlinearities in the evolution of the system (Platzer et al. 2021). The proposed methodology is relevant for various reasons. Firstly, it is written in a Bayesian framework (as BMA and kriging strategies), where data-driven forecasts are considered as Gaussian *a priori* information which are sequentially updated thanks to observations (which are also considered as Gaussian). This provides accurate and reliable initialization for the next forecast. The difference between the data-driven forecasts and the observations allows the computation of the innovation likelihood (e.g., Carrassi et al. 2017), which is used locally in time as metric of model performance. Here also, the scores take into account the model dependency as in the formalism of ClimWIP (Knutti et al. 2017). We then obtain weights that we test by applying them to long-term simulation climatologies.

This study aims to develop a weighting methodology. As a first step, we use numerical simulations of an idealized chaotic deterministic model representing the centennial-to millennial dynamic of Atlantic Meridional Overturning Circulation (AMOC). The proposed alternative approach is compared to more classic approaches, such as model democracy, climatology, or single best model.

The study is organized as follows. Section 2 details the methodological framework for obtaining model weights and applying them. Section 3 describes the numerical model and the data used to implement the methodology. Section 4 is devoted to the numerical results and Sect. 5 concludes this study and presents some perspectives for future research.

2 Methodology

This section first describes the framework to measure the ability of a single dynamical model to fit a set of noisy observations. After applying it to an ensemble of competing models, the second part is devoted to the strategies for computing model weights and their application.

2.1 Evaluating the local performance of a dynamical model

Here, we evaluate the model performance based on its short-term dynamics. By synchronising model forecasts to available observations, Bayesian data assimilation (DA) is a suitable framework for addressing this. DA, like Kalman filtering strategies, is commonly used to improve the estimation of the latent unknown state of a system by sequentially including the available observations (Carrassi et al. 2018). The associated state-space model is expressed as

$$\mathbf{x}_k = \mathcal{M}\left(\mathbf{x}_{k-1}\right) + \boldsymbol{\eta}_k, \tag{1a}$$

$$\mathbf{y}_k = \mathcal{H}\left(\mathbf{x}_k\right) + \boldsymbol{\epsilon}_k, \tag{1b}$$

where k is the time index, $\mathbf{x}_k$ is the true state, $\mathcal{M}$ is the model propagator, $\boldsymbol{\eta}_k$ is the model error, $\mathbf{y}_k$ is the noisy observations, $\mathcal{H}$ is the observational operator, and $\boldsymbol{\epsilon}_k$ is the observation error. Equation (1a) represents the model relation between the true state and the previous state given an additive model error. Equation (1b) makes the relationship between the observations to the true state with an additive observational error.

In the classic Kalman filter, an assimilation cycle aims to sequentially update a Gaussian forecast distribution (generated by the model equations) with the information provided by the observations (also Gaussian), when available. The posterior analysis distribution is a more accurate and reliable estimation of the latent state by satisfying robust statistical properties (i.e., best linear unbiased estimator). The ensemble Kalman filter (EnKF) is preferred when dealing with nonlinear systems (Evensen 2003). At each time step, the two first moments of the Gaussian forecast distribution are approximated empirically using a Monte Carlo approach. To do this, N forecast members are generated from the same model equations, each one initialized from a sample of the previous analysis distribution. This is expressed as

$$\mathbf{x}^f_{(j),k+1} = \mathcal{M}\left(\mathbf{x}^a_{(j),k}\right) + \boldsymbol{\eta}_{(j),k+1} \tag{2}$$

where j is the member sample index, $\mathbf{x}^f_{(j),k+1}$ is the j^{th} forecast member at time $k+1$, $\mathbf{x}^a_{(j),k}$ is the j^{th} analysis member at the previous time k and $\boldsymbol{\eta}_{(j),k+1}$ is the model error assumed to be Gaussian. In Eq. (2), $\mathbf{x}^a_{(j),k}$ is calculated as

$$\mathbf{x}^a_{(j),k} = \mathbf{x}^f_{(j),k} + \mathbf{K}_k\left(\mathbf{y}_k - \mathcal{H}(\mathbf{x}^f_{(j),k}) + \boldsymbol{\epsilon}_{(j),k}\right) \tag{3}$$

where $\boldsymbol{\epsilon}_{(j),k}$ is the observation error drawn from a random sample from the multivariate Gaussian with the 0-mean vector and an error covariance matrix $\mathbf{R}$. The Kalman gain $\mathbf{K}_k$ in Eq. (3) is defined as

$$\mathbf{K}_k = \mathbf{P}^f_k\mathbf{H}^T\left(\mathbf{H}\mathbf{P}^f_k\mathbf{H}^T + \mathbf{R}\right)^{-1}, \tag{4}$$

where $\mathbf{P}^f_k$ is the empirical covariance matrix of the forecast distribution at time k calculated using the EnKF j members, $\mathbf{H}$ represents the tangent linear of the observation operator $\mathcal{H}$.

DA allows the model performance evaluation by measuring the consistency between the model forecasts and the available observations. Directly computed in a DA cycle, the Contextual Model Evidence (CME) is defined as the log-likelihood of observations at a given time for a model, taking into account the forecast state as prior information (Carrassi et al. 2017). Larger CME is obtained for more consistency between the model forecast and observations. Under Gaussian assumption included in the EnKF cycle, CME is approximated as follows

$$\mathrm{CME}_k(\mathbf{y};\mathcal{M}) = -\frac{1}{2}\left(\mathbf{y}_k - \mathcal{H}(\overline{\mathbf{x}}^f_k)\right)\left(\mathbf{H}\mathbf{P}^f_k\mathbf{H}^T + \mathbf{R}\right)^{-1}\left(\mathbf{y}_k - \mathcal{H}(\overline{\mathbf{x}}^f_k)\right)^T - \frac{1}{2}\ln\left(\det\left(\mathbf{H}\mathbf{P}^f_k\mathbf{H}^T + \mathbf{R}\right)\right) - \frac{r}{2}\ln(2\pi), \tag{5}$$

where $\overline{\mathbf{x}}_k^f$ is the empirical mean of the ensemble forecasts and r is the number of available observations. While RMSE is a L^2 metric between the prior forecast state and the observation, similarly the CME can be seen as a distance between the same two fields (Metref et al. 2019). The advantage of CME is that it takes into account reliability information from both forecasts and observations (via covariance matrices) in addition to accuracy information (i.e., the difference between mean states).

In principle, DA formalism requires access to the numerical model to propagate the state of the system in order to obtain the forecasts. When performing EnKF, a large number of model forecasts are generated, representing a significant computational cost. An efficient and flexible data-driven alternative aims at combining DA with a statistical forecasting strategy based on analogs (Lorenz 1969). In Analog Data Assimilation (AnDA), existing long-term simulations are used instead of performing new sequential models simulations (Tandeo et al. 2015; Lguensat et al. 2017). The simulations are decomposed as a catalog of short-term trajectories (i.e., pairs of analog states and their successors) sampling the short-term model dynamics. Hence, using an appropriate distance, the analog states closest to the assimilated state are selected from the catalog. Their successors are combined using a statistical operator to obtain probabilistic forecasts. Hence, AnDA corresponds to perform a classic DA process, but replacing in Eq. (1a) the model propagator by its analog-based approximation. This reads:

$$\mathbf{x}_k = \widehat{\mathcal{M}}(\mathbf{x}_{k-1}) + \boldsymbol{\eta}_k, \tag{6}$$

where $\widehat{\mathcal{M}}$ refers to the analog-based forecasting propagator.

In practice, it consists in fitting a local linear regression (Cleveland and Devlin 1988) relating the selected analogs to their successors, which has proven to be robust (Zhen et al. 2020; Platzer et al. 2021). The regression is able to learn the time-local dynamic of the model including nonlinearities since the linear adjustement is applied locally in the state space. By projecting the current state with the regression, statistical approximated forecasts are obtained. The forecast takes into account the regression uncertainty (assumed Gaussian) which is a robust estimate of the model error. When the catalog is large enough (to approximately cover the full state space), it has been shown that AnDA performs as well as DA (Lguensat et al. 2017). Ruiz et al. 2022 have recently shown, using different idealized dynamical models, that CME can be robustly estimated using AnDA. The combination of AnDA with CME is the basis of the current study.

2.2 Strategies for model weighting-average

The methodology described above is applied to an ensemble of models in order to individually measure their forecast skills given a set of observations. For each model, CME time series are obtained and used to derive individual weights. The weights are then applied to provide a model-average representative of the observations.

2.2.1 Calculating CME-based weights

The CME time series can be processed in various ways to obtain a single scalar (i.e., the score) used to derive the model weight. In this study, three CME scores are defined, together with three benchmark scores (not using the CME). Hence, we define $w^{(i)}_{\text{score-type}}$ as the score assigned to model $\mathcal{M}^{(i)}$. For simplicity, the scores are here described prior to their normalization, which leads to the model weights.

(a) The model democracy (Knutti 2010) gives the same weight to all models, regardless of model performance. It reads:

$$w^{(i)}_{\text{democracy}} = 1. \tag{7a}$$

(b) The climatological score is based on the comparison of model and observation distributions. It measures the minimum cumulative value of the two histograms, highlighting the common area between both distributions. A score close to 0 denotes a model that poorly simulates the observed distribution, while a score of 1 is obtained by a model with a perfect climatological distribution (Perkins et al. 2007). It reads:

$$w^{(i)}_{\text{climato}} = \sum_{j=1}^{n} \text{minimum}(Y(j), M^{(i)}(j)), \tag{7b}$$

where Y and $M^{(i)}$ are the normalized histograms (such that their respective sum equals 1) of the observations and the model simulations, respectively, j is the index of histogram equal-width bins, and n is their total number. Note that in the current study, single-variable observations are used but the score could be adapted using multivariate distributions.

(c) The single best model score assigns 1 to the model showing the best climatological score and 0 to the others. Opposite to the democracy score, it only takes into account the best-performing model over the whole set of observations (Tebaldi and Knutti 2007). It reads:

$$w^{(i)}_{\text{single}} = \begin{cases} 1, & \text{if } w^{(i)}_{\text{climato}} > w^{(j)}_{\text{climato}} \ \forall i \neq j; \\ 0, & \text{otherwise.} \end{cases} \tag{7c}$$

(d) The CME-ClimWIP score is based on the ClimWIP approach (Knutti et al. 2017). Our adaptation reads:

$$w^{(i)}_{\text{CME-ClimWIP}} = \frac{e^{\frac{D_i}{\sigma_D}}}{1 + \sum_{j\neq i}^{L} e^{\frac{S_{ij}}{\sigma_S}}}, \tag{7d}$$

where D_i measures the intra-model performance with its associated shape parameter σ_D and S_{ij} measures the inter-model dependency with its associated shape parameter σ_S. D_i and S_{ij} are usually expressed with Euclidian distance (Knutti et al. 2017). Here, we evaluate D_i such as $D_i = \sum_{k=0}^{K} \text{CME}_k(\mathbf{y}; \mathcal{M}^{(i)})$, where K is the number of time step and S_{ij} as the sum of the defined inter-model such that $S_{ij} = \sum_{k=0}^{K} \text{CME}_k(\mathcal{M}^{(i)}; \mathcal{M}^{(j)})_{j\neq i}$. The CME-ClimWIP score is higher if model forecasts match observations and if they are sufficiently different from other-model forecasts. The two parts of Eq. (7d) are balanced thanks to the use of the parameters σ_D and σ_S. When they are appropriately set within the score, a few top-performing models are selected. Notably, higher values for these parameters bring the score closer to the model democracy while when σ_D=1 and σ_S tends to ∞ the BMA expression is retrieved (see Appendix A for more details on the link between the CME-ClimWIP score and the BMA framework). Note that σ_D and σ_S are determined according to the specific experiments model tests following Knutti et al. 2017 and Lorenz et al. 2018.

(e) The CME best punctual model score exploits the time-local performance provided by the CME. It assigns a local-value of 1 to the best model at time k and 0 to the other models. The score reads:

$$w^{(i)}_{\text{punctual}} = \frac{1}{K}\sum_{k=0}^{K} \mathbb{1}_{\{\text{CME}_k(\mathbf{y};\mathcal{M}^{(i)}) > \text{CME}_k(\mathbf{y};\mathcal{M}^{(j)})\}_{i \neq j}}, \tag{7e}$$

where $\mathbb{1}$ is the indicator function assigning 1 when the condition $\{\text{CME}_k(\mathbf{y};\mathcal{M}^{(i)}) > \text{CME}_k(\mathbf{y};\mathcal{M}^{(j)})\}_{i \neq j}$ is satisfied and 0 otherwise. This score captures the good performance of model in specific regions of the phase space, despite not necessary being optimal in other regions. For example, extreme values are often only well captured by a few specific models. By retaining the information from only one model at each time k, the score bypasses the need to identify temporal inter-model similarities.

(f) The CME best persistent model score $w^{(i)}_{\text{persistent}}$ is derived from the previous score by overweighting models that are the best on several consecutive states. The score is linearly proportional to the cumulative number of consecutive time steps where the model is the best (expression not shown here). This score emphasises consistently good local dynamics.

2.2.2 Assessing the performance of weighting-averages

Here, the methodology consists in two successive stages (Fig. 1). The first step corresponds to a training stage where AnDA is applied to compute the CME time series for each model. The second step is a testing stage in which the CME are used, to compute the weights (i.e, normalized scores) as described in the previous subsection. Applying the weights to individual model distributions leads to a reconstructed distribution. To test the skill of this reconstructed distribution, its overlap with the true distribution is computed (in the same way as the formalism of the climatological score $w^{(i)}_{\text{climato}}$, see Eq. (7b)). Comparison between the overlaps obtained using different scores allows us to compare and then rank the weighting strategies.

3 Experimental setup

3.1 Idealized AMOC model

The study is based on an idealized autonomous low-order deterministic model of the AMOC able of reproducing its millennial variability within a chaotic dynamics (Sévellec and Fedorov 2014; Sévellec and Fedorov 2015). The model derives from the long history of salinity loop-models (originally proposed by Welander 1957; Welander 1965; Welander 1967). Its formalism with three equations (Dewar and Huang 1995; Dewar and Huang 1996; Huang and Dewar 1996) reads:

$$\frac{\mathrm{d}\omega}{\mathrm{d}t} = -\lambda\omega - \epsilon\beta S_{NS}, \tag{8a}$$

$$\frac{\mathrm{d}S_{BT}}{\mathrm{d}t} = (\Omega_0 + \omega)\, S_{NS} - K S_{BT} + \frac{F_0 S_0}{h}, \tag{8b}$$

$$\frac{\mathrm{d}S_{NS}}{\mathrm{d}t} = -(\Omega_0 + \omega)\, S_{NS} - K S_{NS}, \tag{8c}$$

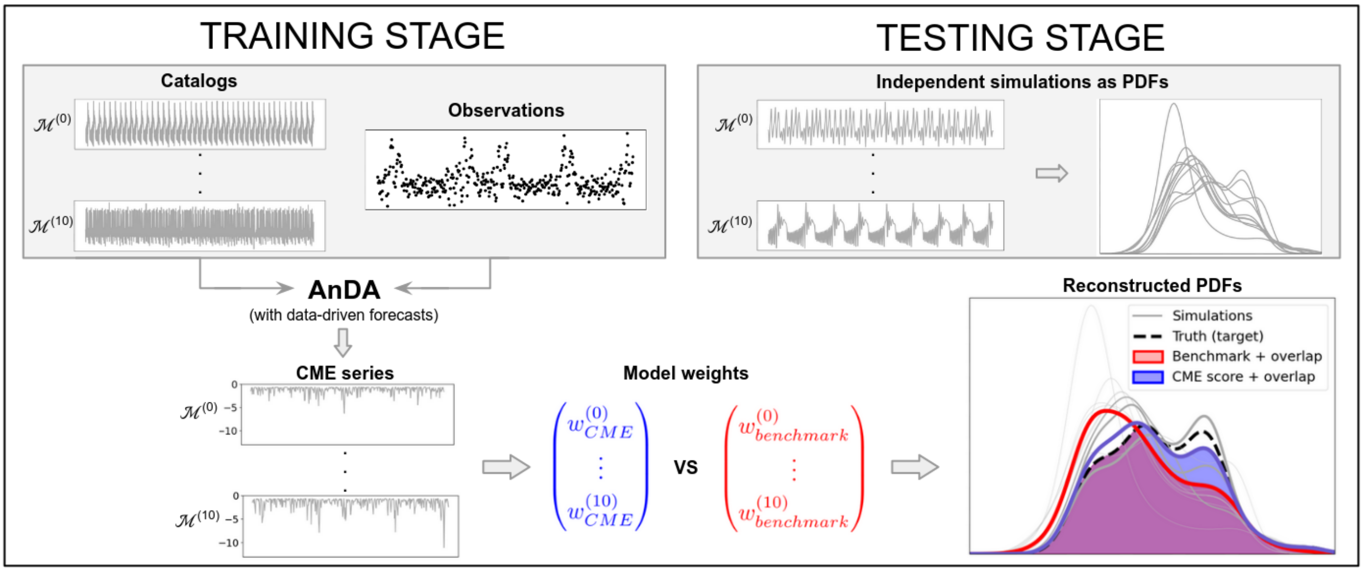


Figure 1. Schematic of the two-stages methodology for one experiment using 11 models. Left panel: training stage performing AnDA with the 11 catalogs and a single set of observations leading to CME computation. Bottom center: from the CME-based score and benchmark score model weights are obtained (blue and red, respectively). Right panel: by applying the weights to individual model distributions (grey), a single reconstructed distribution associated with each score is obtained (blue an red) which is compared to the truth (dashed black). The quality of the reconstruction is assessed using a distribution overlap (blue and red shadings).

where t is the time, ω is the variable component of the overturning circulation strength, S_{BT} and S_{NS} are the bottom-top salinity gradients and the north-south salinity gradients, respectively, Ω_0 is the steady part of the overturning circulation strength which acknowledges the impact of constant temperature and surface winds, ϵ is the buoyancy torque, λ is a linear friction, β is the haline contraction coefficient, K is a linear damping and F_0S_0/h is the surface salt flux (where F_0 is the freshwater flux intensity, S_0 is a reference salinity, and h is the loop thickness i.e., the depth of the level of no motion).

Equation (8a) describes the momentum balance. Here, the rate of change of the ocean overturning strength is driven by the meridional salinity gradients accounting for the buoyancy torque and a linear friction. Equations (8b) and (8c) describe the evolution of the salinity gradients driven by advection, a linear damping representative of diffusion, and the surface salt flux.

Numerical integrations are done using a 4th order Runge-Kutta method with a time step of 1 year. All reference parameters are set following Sévellec and Fedorov 2014 (which we refer interested reader to). Their values are $h = 1000$ m, $\Omega_0 = -2.5 \times 10^{-2}$ yr^{-1}, $F_0 = 1$m yr^{-1}, $S_0 = 35$ psu, $\lambda = 10^{-2}$ yr^{-1}, $\epsilon = 0.35$ yr^{-2}, $K = 10^{-4}$ yr^{-1} and $\beta = 7 \times 10^{-4}$ psu^{-1}.

3.2 Ensemble of AMOC model versions

Various versions of the AMOC model are obtained by perturbing some parameters in the equations: Ω_0, K and $\frac{\lambda}{\epsilon}$, while other remain fixed. All possible combinations between the reference values of the parameters, twice of them, and half of them

https://doi.org/10.5194/egusphere-2023-2649
Preprint. Discussion started: 1 December 2023

lead to twenty-seven distinct parameterizations. Eleven are selected by ignoring the versions describing singular dynamics (e.g., with strong instabilities, constant over time, or presenting periodic behaviour). Thus, for the same equations, the 11 sets of parameters induces their own dynamics. These dynamics show similarities and differences that can be visualised by the computation of the trajectories in the phase space (Fig. 2).

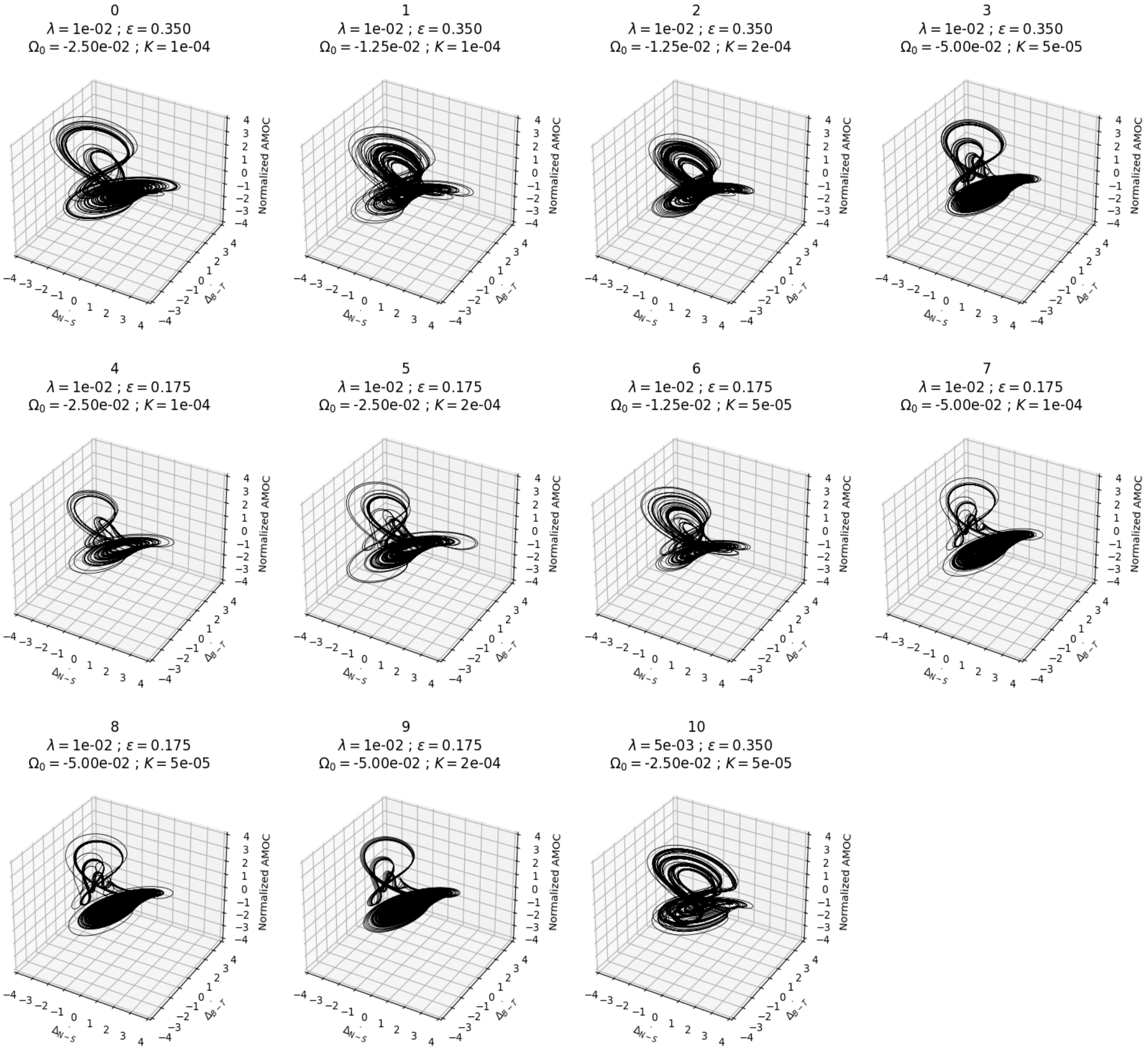


Figure 2. Chaotic trajectories in the phase space of the eleven perturbed versions of the 3-variable (normalized here) AMOC model (Sévellec and Fedorov 2014). The values of the perturbed parameters λ, ϵ, Ω_0 and K used to generate the eleven model versions are indicated in the title. The attractor labelled 0 (top left) corresponds to the model with the reference set of parameters.

https://doi.org/10.5194/egusphere-2023-2649
Preprint. Discussion started: 1 December 2023

3.3 Synthetic data

A model-as-truth experiment framework is set up (e.g., Herger et al. 2018), in which one model is used to generate noisy pseudo-observations. Here, only ω is observed and will be used as the variable of interest. To efficiently select the analogs, the three variables are normalized (here rescaled). In particular, the normalized values of the original variable ω are associated with the new variable called "normalized AMOC" (Fig. 3). In all the experiments, the covariance of the observations for the normalized AMOC is fixed to $\mathbf{R} = 0.5\mathbf{I}_3$. The AnDA framework is applied over a period of K=8000 years with an assimilation step of 20 years, thus resulting in 400 assimilation cycles. In each AnDA cycle, the forecast distributions are estimated by selecting 10000 analogs (which avoids computational divergence for certain attractors with highly nonlinear dynamics in the set) from the associated catalog of size 400 000 years, for each of the 200 EnKF-members. The choice of the assimilation step allows the forecasts to be sufficiently differentiated, so that the competing models can be distinguished locally over time when using the CME (sensitivity experiments not shown).

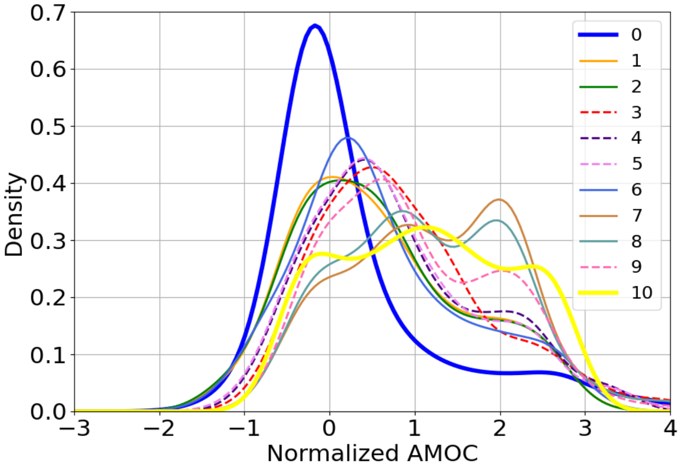


Figure 3. Distributions of normalized AMOC for the eleven model versions

4 Experimental results

230 The different methods are evaluated using two protocols. In the first one, we consider a "perfect experiment", where the model used to generate the pseudo-observations is also included in the catalog. In the second one, the model is excluded from the catalog and we talk about "imperfect experiment". For both perfect and imperfect cases, we first describe an illustrative experiment when model 8 is used as pseudo-observations, and then summarized the results for each model alternately used to generate the pseudo-observations following a Leave-One-Out experiments design.

https://doi.org/10.5194/egusphere-2023-2649
Preprint. Discussion started: 1 December 2023

4.1 Perfect model experiment

The goal of the perfect model approach is to measure the ability of the three CME-based scores to retrieve the correct model (by giving it a predominant weight) from a pool of catalogs, including the true one. The skills of these scores are then compared with the skills of the three benchmarks. In all the perfect model experiments of this study, the CME-ClimWIP score is computed using the optimal values of σ_D=0.3 and σ_S=0.4 which maximize the overall overlap scores in all the experiments combined (following Knutti et al. 2017 and Lorenz et al. 2018).

4.1.1 Experiment with model 8

AnDA is performed on the eleven models with the same pseudo-observations from model 8 and the CME is computed for each assimilation cycle. The resulting CME series for each individual model varies significantly over time and differs locally from model to model (Fig. 4). CME of models 0, 1, 2, 6, and 10 is more than 75% lower than -1 (Fig. 5), showing their inefficiency to replicate model 8 system dynamics. The CME distributions of models 3, 7, and 9 show a narrower distribution with 75% values greater than -1. This means they are both accurate and reliable. As expected, the model with the highest and more consistent CME is model 8. This demonstrates the ability of the CME to retrieve the correct model from a range of models.

The climatologies-based score produces almost uniform weights close to model democracy (Table 1). Indeed, the model climatological distributions are very comparable. This impacts the reliability of selecting a single model following the best model score. Beyond that, the observation error leads the climatologies-based score to misidentify model 9 instead of model 8. The CME-ClimWIP score assigns its highest weight to the correct model 8, although models 7 and 9 have close weights. Regarding the two other CME-based scores, although the weight associated to model 8 is high, these scores give the highest weight to model 9. We conclude that the time selection characteristics of them make both scores more sensitive to observation errors.

	0	1	2	3	4	5	6	7	8	9	10
Model democracy	9.1	9.1	9.1	9.1	9.1	9.1	9.1	9.1	9.1	9.1	9.1
Climatologies	6.2	9	9.1	9.8	9.4	9.5	8.7	9.3	9.7	**10**	9.2
Best single model	0	0	0	0	0	0	0	0	0	**100**	0
CME ClimWIP	0.1	0	0	16.2	0.3	0.5	0.1	30.4	**31.2**	21.2	0
CME best punctual	4	2.3	2	10	8.3	6.5	4.8	12.3	16.8	**24**	9.3
CME best persistent	3.3	1.9	1.7	8.5	7.3	6	5	12.5	17.5	**28.1**	8.1

Table 1. Perfect model results for model 8 as the pseudo-observation. For each score (in row): weights (in %) associated with the eleven models. The sum of the weights per row is 100. Values in bold highlight the model with the highest weight for each score.

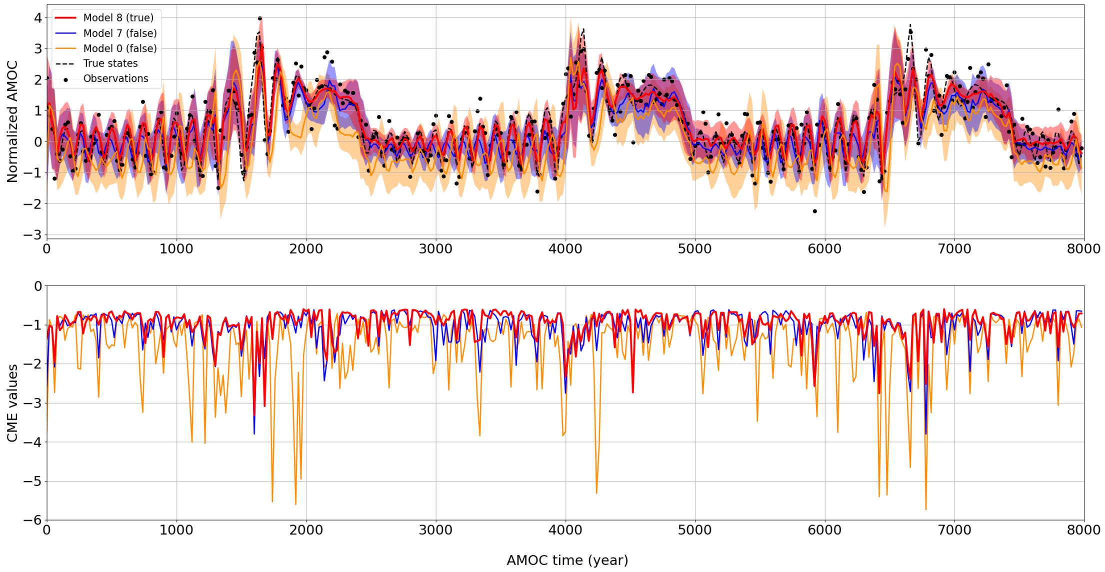


Figure 4. AnDA results on 400 cycles using model 8 to generate the pseudo-observations: examples of three assimilated models among the eleven. Top panel: time series of normalized AMOC with forecasts mean states and 90% prediction interval (in shades). The red series denotes the true model (i.e., model 8), the orange and blue ones refer to models 0 and 7, respectively. The black dots represent the noisy observations generated from the true states denoted by the dashed black line. Bottom panel: associated CME time series for each model. Close to zero values of the CME indicate better forecast distribution.

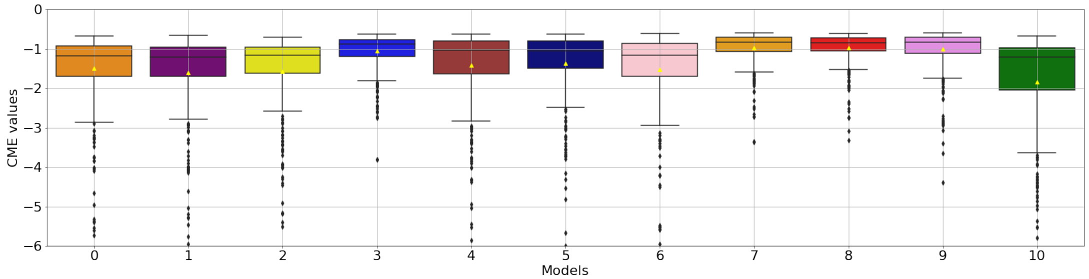


Figure 5. CME distributions from AnDA results using model 8 to generate the pseudo-observations. CME distribution are on the y-axis and model number (from index 0 to 10) on the x-axis. The boxes span the 25th, median, and 75th percentiles of the distributions, while the whiskers show the minimum and maximum values (excluding the outliers denoted with the black diamonds). Mean is indicated with the small yellow triangle.

4.1.2 Overall eleven experiments

By varying the model used to generate the pseudo-observations in the eleven AnDA experiments, we can robustly assess the ability of the six scores to efficiently retrieve the correct model (Table 2). As emphasized in the specific experiment for model 8, the scores based on climatology do not allow a clear discrimination between models. Hence the climatological and the best-single-model only retrieved the correct model twice over the eleven experiments (i.e., experiments with models 1 and 10). The CME-based scores perform better, attributing their highest weight to the correct model 8 times for CME-ClimWIP and both 6 times for CME best punctual and CME best persistent. As for model 8 experiment, in most of the unsuccessful experiments of the three CME scores, the correct model still has a weight close to the weight of the selected one (not shown here).

	0	1	2	3	4	5	6	7	8	9	10	# success
Model democracy	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0/11
Climatologies	6	**1**	1	5	5	3	1	3	9	8	**10**	2/11
Best single model	6	**1**	1	5	5	3	1	3	9	8	**10**	2/11
CME ClimWIP	**0**	**1**	**2**	**3**	5	**5**	**6**	8	**8**	8	**10**	8/11
CME best punctual	**0**	10	10	**3**	5	**5**	**6**	9	9	**9**	**10**	6/11
CME best persistent	**0**	10	10	**3**	5	**5**	**6**	9	9	**9**	**10**	6/11

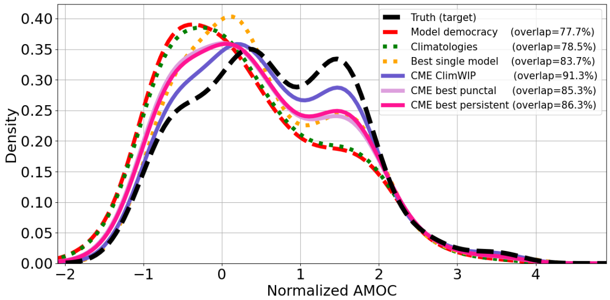
Table 2. Summarised perfect model results of the Leave-One-Out model-as-truth experiments. For each column representing an experiment (i.e., the model index used to generate pseudo-observations), the index of the model with the highest weight is specified for each score in the row. The indices in bold show when the true model is recovered. The last column summarizes how frequently the correct model is identified among the candidate models across the 11 experiments for the five scores. Note that 'N/A' is indicated for model democracy since the score prevents differentiation between models.

4.2 Imperfect model experiment

The imperfect model approach aims to reconstruct the statistical properties of the distribution of a missing model using distinct models. In all the imperfect experiments of this study, the CME-ClimWIP score is computed using the optimal values of $\sigma_D = 0.5$ and $\sigma_S = 0.7$ which maximize the overlap scores in all the experiments combined (following Knutti et al. 2017 and Lorenz et al. 2018).

4.2.1 Experiment with model 8

The model democracy approach yields the worst reconstruction (Fig. 6, left panel), stressing that, in this case, the observations information should be included in the weight calculation. The climatology-based score slighty improves the reconstruction provided by the model democracy (+0.8% compared to model democracy). By selecting model 9, the best single-model score improve by 6% the reconstruction of the true model 8 distribution, when compared to the model democracy. The reconstructed distributions associated with the three strategies (i.e., model democracy, climatology, best single model) are all positively skewed, underestimating the true normalized AMOC (Fig. 6, left panel). The three CME-based scores outperform the three benchmark ones, with greater overlaps of their reconstruction with the truth. In particular, they all reconstruct better statistics for high positive values of normalized AMOC. For CME-ClimWIP, three models contribute the most to the reconstruction of the true missing model: model 3 by 18.1%, model 7 by 49.1%, and model 9 by 31.7%. To a lesser extent, the same three models are also selected for the best punctual and best persistent scores, but still leaves a greater contribution from other seven models. This suggests that the information provided by these seven models is negligible, as the two last CME strategies have lower overlapping scores than the one obtained using CME-ClimWIP.



	0	1	2	3	4	5	6	7	8	9	10
Model democracy	10	10	10	10	10	10	10	10		10	10
Climatologies	6.8	10	10	**10.9**	**10.5**	**10.6**	9.6	**10.3**		**11.1**	**10.2**
Best single model	0	0	0	0	0	0	0	0		**100**	0
CME ClimWIP	0.1	0	0.1	**18.1**	0.3	0.5	0.1	**49.1**		**31.7**	0
CME best punctual	4	2.2	2.2	**10.5**	8.3	6.4	4.8	**26.8**		**25.5**	9.3
CME best persistent	3.2	1.9	1.9	9.1	7.1	5.8	4.9	**29.7**		**28.6**	7.8

Figure 6. Imperfect model results of model 8 experiment (by excluding it). Left panel: reconstructed distribution of normalized circulation strength for the six scores with different colors. They are compared with the true distribution in dash black. The overlap score (in %) for each score is expressed in the legend in brackets. Right panel: table of model weights (in %) for each score (the sum of the weights per row is 100). Values in bold highlight the weights greater than the model democracy weight (>10%) which correspond to the models contributing the most to the distribution reconstruction.

4.2.2 Overall eleven experiments

When the eleven models are reconstructed independently in the imperfect framework, the three CME-based scores produce better reconstructions on average than the democracy and climatologies scores (Fig. 7). Similar to the specific model 8 reconstruction, the climatologies comparison score does not provide meaningful distinctions between the models. This results in high variations in the reconstruction score when using the best single model across all experiments, further highlighting its unreliability. For instance, while the score effectively captures most of the distribution when reconstructing models 1 and 2 (due to their highly similar dynamics), for other experiments, it yields lower reconstruction performance.

The reconstruction performances in the eleven experiments show that CME-ClimWIP greatly outperforms the model democracy for 7 of 11 model reconstructions. As for model 8 reconstruction, CME-ClimWIP gives higher weights to a small subset of models, making it well suited to reconstruct distributions that share similarities with some others in the ensemble.

On the other hand, CME-ClimWIP is close but less useful than democracy (and climatologies) for reconstructing models 3, 5, 9, and 10. Here, despite the six scores having lost performance, uniform weights are appropriate for good reconstruction. A typical example of such case is the reconstruction of model 10 whose distribution is quite symmetric, whereas all the other models have asymmetric distribution. For model 3, 5, 9, and 10 experiments, CME best punctual and persistent give better results than CME-ClimWIP. It is worth noting that these two scores also have greater reliability across all eleven experiments than CME-ClimWIP, since their reconstruction performance remains consistently superior to the model democracy score with minimal variation between all the experiments.

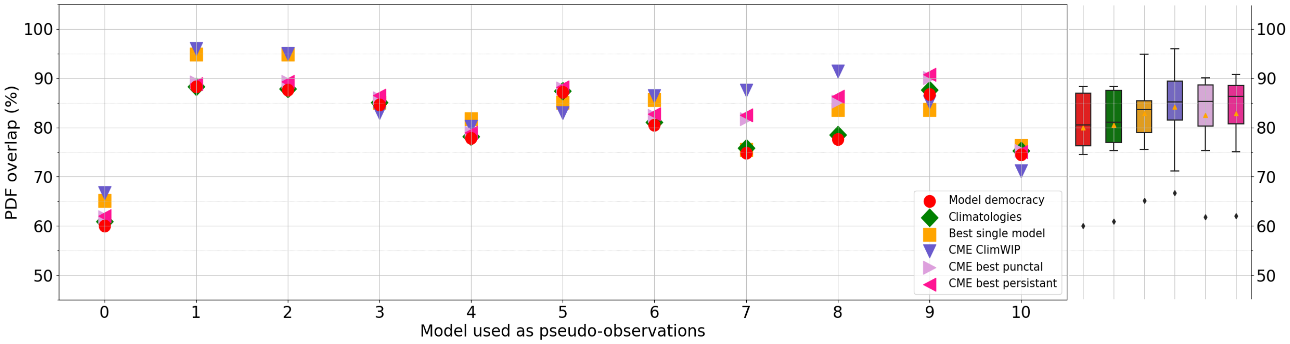


Figure 7. Imperfect model results (i.e., excluding the true model) of the Leave-One-Out model-as-truth experiments. Left panel: the model index used as pseudo-observations is on x-axis, whereas the overlap scores (in %) between the reconstructed PDF and the true PDF read in y-axis. Each colored marker corresponds to a score strategy as denoted in the legend. Right panel: boxplots spanning for each score (keeping the same colour code) the 25th, median, and 75th percentiles of the distributions of overlap scores for all the model-as-truth experiments combined. The whiskers show the minimum and maximum values excluding the outliers denoted with the black diamonds. Small yellow triangle indicate the means.

4.2.3 Relative influence of the CME-ClimWIP score components

As expressed in (7d), the CME-ClimWIP score is composed of the performance and dependence of the models, both expressed using the CME. The CME at time k depends on the innovation (i.e., the difference between $\overline{\mathbf{x}}_k^f$ and $\mathbf{y}_k$) and the covariance error matrices $\mathbf{P_k^f}$ and $\mathbf{R}$. The degree of contribution of the performance and independence is controlled by two tuned shape parameters: σ_D and σ_S. Hence, to assess the importance of each component in the reconstruction performance, various versions of the CME-ClimWIP score are tested, by taking into account or not the influence of each component (Table 3).

The degree of dependence between models does not contribute significantly to the reconstruction performance (Table 3). Since the eleven model versions are based on the same set of equations and the three variables are normalized, the model-model redundancy is significant. This implies that the inter-model dependency (i.e., the denominator in Eq. (7d)), is too similar across the models to contribute notably to differentiate them. The shape parameter (σ_D) has the main impact on the CME-ClimWIP by driving the score performance. The reconstruction performance of the CME-ClimWIP version without tuning the shape parameters is more than 10% lower to the reference performance. The score computed without information from $\mathbf{P}^f$ and $\mathbf{R}$ has lower performance than the reference, confirming the usefulness of CME (including the covariances in its formulation) over a metric only based on Euclidian distance. Finally, when both covariances and shape parameters are excluded from the score, the performance lost is even more important.

CME-ClimWIP alternative/simplified versions	model 8 experiment (%)	All experiments mean [mean-std , mean+std] (%)
CME-ClimWIP reference	91.30	84.15 $[75.09, 93.29]$
CME-ClimWIP without codependency part (i.e., $\sigma_S \to +\infty$)	90.93	84.11 $[75.09, 93.13]$
CME-ClimWIP with σ_D=σ_S=1	80.32	81.19 $[72.81, 89.58]$
Score without $\mathbf{P}^f$ and $\mathbf{R}$ informations	89.45	83.68 $[74.61, 92.75]$
Score without $\mathbf{P}^f$ or $\mathbf{R}$ and with σ_D=σ_S=1	79.24	80.84 $[72.47, 89.21]$

Table 3. Sensitivity of the reconstruction performance (based on overlap of the reconstructed distribution with the truth) to various alternative/simplified versions of the CME-ClimWIP score in the imperfect model approach. The second column shows the overlap associated with model 8 experiment. The third column shows the results of all Leave-One-Out experiments combined

5 Conclusion and perspectives

This study aims at developing a data-driven methodology for weighting models, only based on their ability to represent the dynamics of observations. For this purpose, a set of dynamical models is compared to noisy observations, where model equations are replaced by available simulations. The method combines a machine learning approach (i.e., an analog forecasting method) to estimate the model forecasts in a cost-effective manner (i.e., without the need to re-run the model) and a sequential data assimilation algorithm (i.e., the ensemble Kalman filter). The time-varying performance of the models with respect to observations is evaluated using the contextual model evidence (also known as innovation likelihood), which benefits from DA properties.

To test this methodology, an ensemble of eleven models is obtained by perturbing parameters of an idealized chaotic model of the AMOC. For each model version, the equations are only used to generate large simulations of its three variables. Each version alternately plays the role of truth (i.e., Leave-One-Out experiments) used to construct pseudo-observations of the AMOC component. In the eleven experiments, model weights are extracted from the CME series using different strategies. They take into account the performance of the models with respect to pseudo-observations and the degree of similarity with other models. The method is then assessed by applying the weights to the distributions of long-term model simulations. In this way, we test the extent to which the short-term dynamics of individual models can provide relevant information for reconstructing the statistics of a targeted distribution. Reconstruction performance is measured by the percentage of overlap between the reconstructed and the true distributions. The reconstruction performance associated with the three CME-based scores is compared to three benchmark approaches that do not include DA framework (i.e., model democracy, climatological distributions comparison, and best single model).

The results of the perfect model approach highlight a better performance of CME-based scores in recovering the correct model, compared to benchmarks, suggesting the importance of using DA. The results of the benchmark strategies generally suffer from a lack of discrimination between models to correctly identify the right one. In the context of the imperfect models approach, the scores based on the CME are able to reconstruct the targeted distribution more suitably than the benchmarks using the same partial noisy observations. The results underline that CME-based scores can be seen as a compromise between the "democracy score" and the so-called "dictatorial score" which selects only a single model. Within the CME-based score, the CME-ClimWIP score is relatively closer to "dictatorship". It is suitable for reconstructing distributions sharing similarities with few models in the ensemble and when democracy is not appropriate. On the other hand, the CME best punctual and CME persistent scores are closer to "democracy". By exploiting temporal performance, their weights are more adaptive, which is advantageous for outperforming the already successful democracy in any experiment (when CME-ClimWIP does not succeed).

This study initiates the implementation of the methodological framework whose ultimate goal is to weight a larger ensemble of dynamic models representing more complex systems, such as those provided by the CMIP6 (Eyring et al. 2016). Here, the aim will be to find weights that will be calculated on the performance of climate models (represented by their catalog of simulations) in correctly reproducing current partial observations. Extrapolating the weighted model averages using future climate projection simulations could provide a more accurate and reliable response of the AMOC to anthropogenic forcing. However, some aspects need to be examined due to the differences between a CMIP study and the present one. Firstly, whereas here the simulations are stationary, the CMIP simulations describe a trend and variabilities in response to the time-varying forcing (e.g., solar activity, volcano eruption, and greenhouse gas emissions). Secondly, in this proof of concept study, the full phase space of the idealized model is known, whereas using CMIP simulations, additional variables would need to be investigated to increase the unknown number of dimensions. This could be valuable for obtaining accurate and reliable AnDA reconstructions needed for deriving relevant weights. The diversity of available observations types could be particularly beneficial in this case.

Code and data availability. Python codes are available in the following GitHub open repository: https://github.com/pilebras/AnDA_weight_idealAMOC/tree/main under the GNU license. They include the code for generating the data (using an idealized AMOC model developped in Sévellec and Fedorov 2014), performing the experiments using AnDA (adapted from the code available at https://github.com/ptandeo/AnDA) and obtaining the figures of the paper.

Author contributions. All authors contributed equally to the experimental design. PLB wrote the article and performed the numerical experiments and the analyses. FS, PT, JR and PA helped on the redaction of the paper.

Competing interests. At least one of the (co-)authors is a member of the editorial board of Nonlinear Processes in Geophysics.

Acknowledgements. The authors thank the financial support for the AMIGOS project from the Région Bretagne (ARED fellowship for PLB) and from the ISblue project (Interdisciplinary graduate school for the blue planet, ANR-17-EURE-0015) co-funded by a grant from the French government under the program Investissements d'Avenir. The authors also want to thank Maxime Beauchamp and Noémie Le Carrer for their helpful comments on the study.

References

Abramowitz, Gab, Nadja Herger, Ethan Gutmann, Dorit Hammerling, Reto Knutti, Martin Leduc, Ruth Lorenz, Robert Pincus, and Gavin A Schmidt (2019). "ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing". In: *Earth System Dynamics* 10.1, pp. 91–105.

Amos, Matt, Paul J Young, J Scott Hosking, Jean-François Lamarque, N Luke Abraham, Hideharu Akiyoshi, Alexander T Archibald, Slimane Bekki, Makoto Deushi, Patrick Jöckel, et al. (2020). "Projecting ozone hole recovery using an ensemble of chemistry–climate models weighted by model performance and independence". In: *Atmospheric Chemistry and Physics* 20.16, pp. 9961–9977.

Brunner, Lukas, Ruth Lorenz, Marius Zumwald, and Reto Knutti (2019). "Quantifying uncertainty in European climate projections using combined performance-independence weighting". In: *Environmental Research Letters* 14.12, p. 124010.

Brunner, Lukas, Angeline G Pendergrass, Flavio Lehner, Anna L Merrifield, Ruth Lorenz, and Reto Knutti (2020). "Reduced global warming from CMIP6 projections when weighting models by performance and independence". In: *Earth System Dynamics* 11.4, pp. 995–1012.

Carrassi, Alberto, Marc Bocquet, Laurent Bertino, and Geir Evensen (2018). "Data assimilation in the geosciences: An overview of methods, issues, and perspectives". In: *Wiley Interdisciplinary Reviews: Climate Change* 0.0, e535. DOI: 10.1002/wcc.535. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcc.535. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/wcc.535.

Carrassi, Alberto, Marc Bocquet, Alexis Hannart, and Michael Ghil (2017). "Estimating model evidence using data assimilation". In: *Quarterly Journal of the Royal Meteorological Society* 143.703, pp. 866–880.

Cleveland, William S and Susan J Devlin (1988). "Locally weighted regression: an approach to regression analysis by local fitting". In: *Journal of the American statistical association* 83.403, pp. 596–610.

Dewar and Huang (1995). "Fluid flow in loops driven by freshwater and heat fluxes". In: *Journal of Fluid Mechanics* 297, pp. 153–191.

– (1996). "On the forced flow of salty water in a loop". In: *Physics of Fluids* 8.4, pp. 954–970.

Evensen, Geir (2003). "The ensemble Kalman filter: Theoretical formulation and practical implementation". In: *Ocean dynamics* 53.4, pp. 343–367.

Eyring, Veronika, Sandrine Bony, Gerald A Meehl, Catherine A Senior, Bjorn Stevens, Ronald J Stouffer, and Karl E Taylor (2016). "Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization". In: *Geoscientific Model Development* 9.5, pp. 1937–1958.

Eyring, Veronika, Peter M Cox, Gregory M Flato, Peter J Gleckler, Gab Abramowitz, Peter Caldwell, William D Collins, Bettina K Gier, Alex D Hall, Forrest M Hoffman, et al. (2019). "Taking climate model evaluation to the next level". In: *Nature Climate Change* 9.2, pp. 102–110.

Garthwaite, PH and E Mubwandarikwa (2010). *Selection of weights for weighted model averaging: prior weights for weighted model averaging. Aust N Zeal J Stat 52 (4): 363–382.*

George, Edward I (2010). "Dilution priors: Compensating for model space redundancy". In: *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*. Vol. 6. Institute of Mathematical Statistics, pp. 158–166.

Giorgi, Filippo and Linda O Mearns (2002). "Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the "reliability ensemble averaging"(REA) method". In: *Journal of climate* 15.10, pp. 1141–1158.

Herger, Nadja, Gab Abramowitz, Reto Knutti, Oliver Angélil, Karsten Lehmann, and Benjamin M Sanderson (2018). "Selecting a climate model subset to optimise key ensemble properties". In: *Earth System Dynamics* 9.1, pp. 135–151.

Huang and Dewar (1996). "Haline circulation: Bifurcation and chaos". In: *Journal of physical oceanography* 26.10, pp. 2093–2106.

IPCC (2021). *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Vol. In Press. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. DOI: 10.1017/9781009157896.

Knutti, Reto (2010). *The end of model democracy?*

Knutti, Reto, Christoph Baumberger, and Gertrude Hirsch Hadorn (2019). "Uncertainty quantification using multiple models—Prospects and challenges". In: *Computer Simulation Validation*. Springer, pp. 835–855.

Knutti, Reto, Reinhard Furrer, Claudia Tebaldi, Jan Cermak, and Gerald A Meehl (2010). "Challenges in combining projections from multiple climate models". In: *Journal of Climate* 23.10, pp. 2739–2758.

Knutti, Reto, Jan Sedlácek, Benjamin M Sanderson, Ruth Lorenz, Erich M Fischer, and Veronika Eyring (2017). "A climate model projection weighting scheme accounting for performance and interdependence". In: *Geophysical Research Letters* 44.4, pp. 1909–1918.

Lguensat, Redouane, Pierre Tandeo, Pierre Ailliot, Manuel Pulido, and Ronan Fablet (2017). "The analog data assimilation". In: *Monthly Weather Review* 145.10, pp. 4093–4107.

Lorenz (1969). "Atmospheric predictability as revealed by naturally occurring analogues". In: *Journal of Atmospheric Sciences* 26.4, pp. 636–646.

Lorenz, Nadja Herger, Jan Sedlácek, Veronika Eyring, Erich M Fischer, and Reto Knutti (2018). "Prospects and caveats of weighting climate models for summer maximum temperature projections over North America". In: *Journal of Geophysical Research: Atmospheres* 123.9, pp. 4509–4526.

Merrifield, Anna Louise, Lukas Brunner, Ruth Lorenz, Iselin Medhaug, and Reto Knutti (2020). "An investigation of weighting schemes suitable for incorporating large ensembles into multi-model ensembles". In: *Earth System Dynamics* 11.3, pp. 807–834.

Metref, Sammy, Alexis Hannart, Juan Ruiz, Marc Bocquet, Alberto Carrassi, and Michael Ghil (2019). "Estimating model evidence using ensemble-based data assimilation with localization – The model selection problem". In: *Quarterly Journal of the Royal Meteorological Society* 145.721, pp. 1571–1588. DOI: 10.1002/qj.3513. eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3513. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3513.

Min, Seung-Ki, Daniel Simonis, and Andreas Hense (2007). "Probabilistic climate change predictions applying Bayesian model averaging". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 365.1857, pp. 2103–2116.

Olson, Roman, Soon-Il An, Yanan Fan, and Jason P Evans (2019). "Accounting for skill in trend, variability, and autocorrelation facilitates better multi-model projections: Application to the AMOC and temperature time series". In: *Plos one* 14.4, e0214535.

Olson, Roman, Yanan Fan, and Jason P Evans (2016). "A simple method for Bayesian model averaging of regional climate model projections: Application to southeast Australian temperatures". In: *Geophysical Research Letters* 43.14, pp. 7661–7669.

Perkins, SE, AJ Pitman, NJMcAneney J Holbrook, and J McAneney (2007). "Evaluation of the AR4 climate models' simulated daily maximum temperature, minimum temperature, and precipitation over Australia using probability density functions". In: *Journal of climate* 20.17, pp. 4356–4376.

https://doi.org/10.5194/egusphere-2023-2649
Preprint. Discussion started: 1 December 2023

Platzer, Paul, Pascal Yiou, Philippe Naveau, Pierre Tandeo, Yicun Zhen, Pierre Ailliot, and Jean-François Filipot (2021). "Using local dynamics to explain analog forecasting of chaotic systems". In: *Journal of the Atmospheric Sciences*.

Pulido, Manuel, Pierre Tandeo, Marc Bocquet, Alberto Carrassi, and Magdalena Lucini (2018). "Stochastic parameterization identification using ensemble Kalman filtering combined with maximum likelihood methods". In: *Tellus A: Dynamic Meteorology and Oceanography* 70.1, pp. 1–17.

Raftery, Adrian E, Tilmann Gneiting, Fadoua Balabdaoui, and Michael Polakowski (2005). "Using Bayesian model averaging to calibrate forecast ensembles". In: *Monthly weather review* 133.5, pp. 1155–1174.

Ribes, Aurélien, Julien Boé, Sai¨d Qasmi, Brigitte Dubuisson, Hervé Douville, and Laurent Terray (2022). "An updated assessment of past and future warming over France based on a regional observational constraint". In: *Earth System Dynamics* 13.4, pp. 1397–1415.

Ribes, Aurélien, Sai¨d Qasmi, and Nathan P Gillett (2021). "Making climate projections conditional on historical observations". In: *Science Advances* 7.4, eabc0671.

Ruiz, Juan, Pierre Ailliot, Thi Tuyet Trang Chau, Pierre Le Bras, Valérie Monbet, Florian Sévellec, and Pierre Tandeo (2022). "Analog Data Assimilation for the Selection of Suitable General Circulation Models". In: *Geoscientific Model Development Discussions*, pp. 1–30.

Sanderson, Reto Knutti, and Peter Caldwell (2015). "A representative democracy to reduce interdependency in a multimodel ensemble". In: *Journal of Climate* 28.13, pp. 5171–5194.

Sanderson, Michael Wehner, and Reto Knutti (2017). "Skill and independence weighting for multi-model assessments". In: *Geoscientific Model Development* 10.6, pp. 2379–2395.

Sévellec and Fedorov (2014). "Millennial variability in an idealized ocean model: Predicting the AMOC regime shifts". In: *Journal of Climate* 27, pp. 3551–3564.

– (2015). "Unstable AMOC during glacial intervals and millennial variability: The role of mean sea ice extent". In: *Earth and Planetary Science Letters* 429, pp. 60–68.

Sexton, David MH, James M Murphy, Mat Collins, and Mark J Webb (2012). "Multivariate probabilistic projections using imperfect climate models part I: outline of methodology". In: *Climate dynamics* 38, pp. 2513–2542.

Tandeo, Pierre, Pierre Ailliot, Marc Bocquet, Alberto Carrassi, Takemasa Miyoshi, Manuel Pulido, and Yicun Zhen (2020). "A review of innovation-based methods to jointly estimate model and observation error covariance matrices in ensemble data assimilation". In: *Monthly Weather Review* 148.10, pp. 3973–3994.

Tandeo, Pierre, Pierre Ailliot, Juan Ruiz, Alexis Hannart, Bertrand Chapron, Anne Cuzol, Valérie Monbet, Robert Easton, and Ronan Fablet (2015). "Combining analog method and ensemble data assimilation: application to the Lorenz-63 chaotic system". In: *Machine Learning and Data Mining Approaches to Climate Science: proceedings of the 4th International Workshop on Climate Informatics*. Springer, pp. 3–12.

Tebaldi, Claudia and Reto Knutti (2007). "The use of the multi-model ensemble in probabilistic climate projections". In: *Philosophical transactions of the royal society A: mathematical, physical and engineering sciences* 365.1857, pp. 2053–2075.

Welander, Pierre (1957). "Note on the self-sustained oscillations of a simple thermal system". In: *Tellus* 9.3, pp. 419–420.

– (1965). *Steady and oscillatory motions of a differentially heated fluid loop*. Woods Hole Oceanographic Institution.

– (1967). "On the oscillatory instability of a differentially heated fluid loop". In: *Journal of Fluid Mechanics* 29.1, pp. 17–30.

Zhen, Yicun, Pierre Tandeo, Stéphanie Leroux, Sammy Metref, Thierry Penduff, and Julien Le Sommer (2020). "An adaptive optimal interpolation based on analog forecasting: application to SSH in the Gulf of Mexico". In: *Journal of Atmospheric and Oceanic Technology* 37.9, pp. 1697–1711.

Appendix A: Link between CME-ClimWIP and Bayesian Model Averaging (BMA)

Including CME as a metric in the CME-ClimWIP expression brings the score into the Bayesian Model Averaging (BMA) framework which relies on marginal likelihoods. Specifically, $e^{\sum_{k=0}^{K}\mathrm{CME}_k(\mathbf{y};\mathcal{M}^{(i)})}$, in the numerator of Eq. (7d), represents the approximate marginal likelihood of the observations. This is estimated within the AnDA process using successive innovation likelihoods (i.e., the CME values, e.g., Carrassi et al. 2017; Pulido et al. 2018; Metref et al. 2019; Tandeo et al. 2020). This approach offers two key advantages. Firstly, it avoids the need for calculating a climatological distribution, which can be challenging to estimate accurately and computationally expensive. Secondly, it provides more informative insights into current conditions, including forecast states with their uncertainties. This is especially useful for dealing with nonlinearities in dynamic systems, instead of using globally defined climatological distributions (Metref et al. 2019). Furthermore, the denominator in Eq. (7d) serves as a prior that contains information about inter-model dependency. This provides more valuable insights compared to a uniform prior commonly used in standard BMA approaches (e.g., George 2010; Garthwaite and Mubwandarikwa 2010).