NPG-2023-2649
Authors: Pierre Le Bras, Florian Sevellec, Pierre Tandeo, Juan Ruiz and Pierre Alliot
Discussion started: 01 December, 2023
Title: Selecting and weighting dynamical models using data-driven approaches

Response to the reviewers

In the response below:

- Commentary from the reviewers are in **black bold.**
- Response to the commentaries are in light blue.
- Text drawn from the manuscript is in *italic light blue*.
- Modified or added text within the manuscript is highlighted by *underlining* it.

Reviewer 1

General comments:

This study developed a new approach for selecting and weighting dynamical models within a multi-model ensemble. The new points are the estimation of weights in the data assimilation framework of EnKF and the data-driven approach to forward the evolution of states. These two aspects may have the potential for further applications. In addition, the manuscript was well written and logically organized. However, some major comments have to be addressed before the manuscript can be considered for publication.

We would like to thank the reviewer for the comments that helped improve the clarity of the manuscript.

1. **It seems that the posterior probability density function (PDF) by assimilating observations in the EnKF algorithm is key to determine the weights of each dynamical system. One limitation of this methodology is the assumption that the variables among multi-model ensembles follow a Gaussian distribution. The systematic errors and the non-Gaussianity would make the performance of this algorithm suboptimal. My question is if the nonlinear particle filter used in the same framework potentially enhances the performance of the new approach. It is worth some discussion in the conclusion section.**

We thank the reviewer for this comment.

We have investigated this key question by computing the Gaussianity of the EnKF member distribution. To assess this, we employ the Jarque-Bera statistical test, which evaluates the Gaussianity of a distribution based on its skewness and kurtosis (Jarque and Bera, 1980). To illustrate this test we have used observations derived from model #0 and used analog-catalog derived from model #8 (Fig. R1). The Gaussian assumption holds true for the test across 50.5% of the time steps.

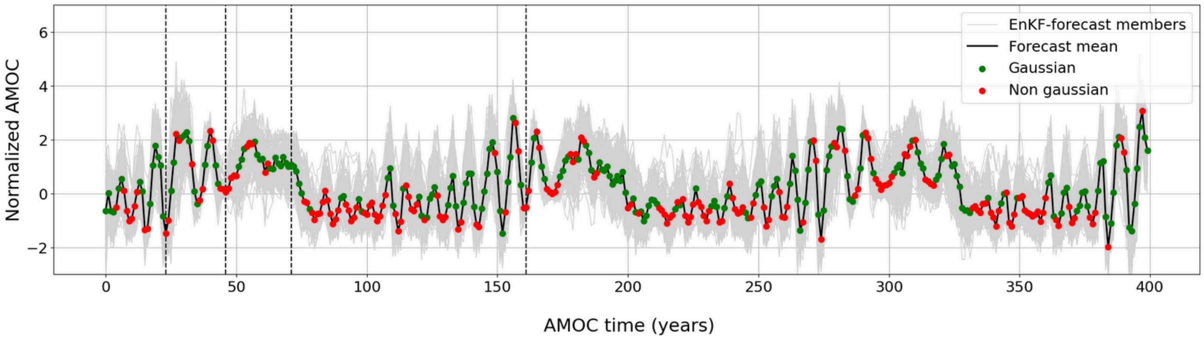


Fig. R1: EnKF-forecast members (grey) and forecast mean (black) of model #0 assimilated with pseudo-observations from model #8. At each time step, a statistical test of Gaussianity is performed (Jarque-Bera test). Significant Gaussian forecast distributions (at 5% significance level) over time are denoted with green dots while the insignificant ones are shown in red.

To further illustrate the Gaussinality we have plotted the ENKF-forecast distribution at a few specific times where the Jarque-Bera statistical test indicated Gaussianity or not (Fig. R2). Our analysis shows that even in cases where the Gaussianity is rejected by the Jarque-Bera criteria, the distributions remain close to Gaussianity.

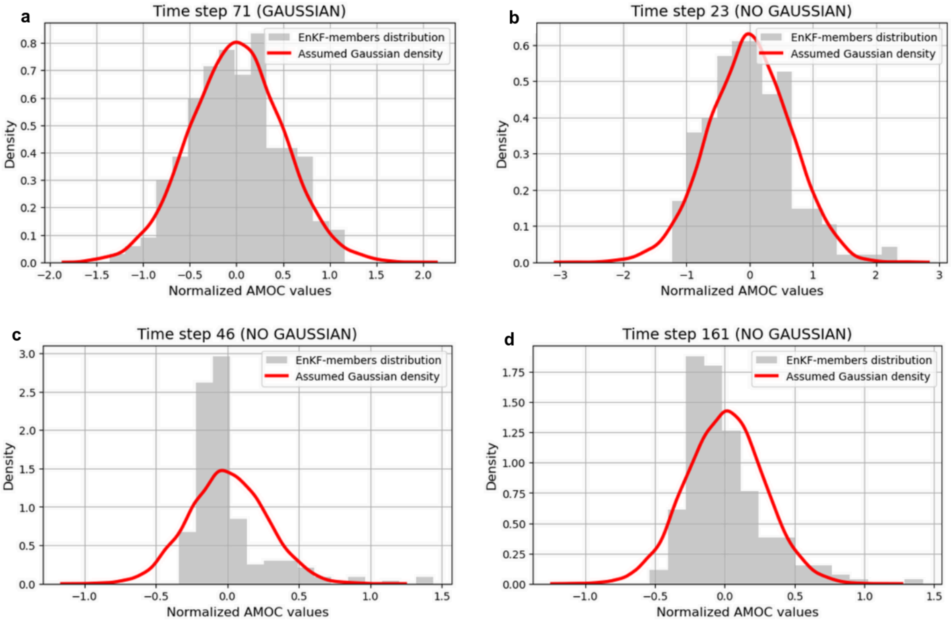


Fig. R2: EnKF-forecast members distributions and associated assumed Gaussian PDF at four extracted time steps from the forecasts series shown in Fig. R1. (a) and (b, c, and d) are cases where the Jarque-Bera test indicates Gaussianity and non-Gaussianity, respectively.

However, to be more general, we agree with the reviewer that non-parametric (e.g., particle filters) should be explored in the future. We have commented on this aspect in the manuscript.

To clarify this aspect, we add the following paragraph in the conclusion section.

Sec. 5, Page 18, Line 382: *“An inherent assumption of our study is the Gaussianity of the EnKF-forecast distributions. Testing its validity through a Jarque-Bera test* (Jarque and Bera, 1980)*, we find that it is valid up to 50.5% of the 400 total forecasts for model #0 and pseudo-observations from model #8. To go beyond our Gaussian approach, non-parametric assimilation method, such as particle filters (Van Leeuwen, 2009), could be implemented. This approach is more suitable where the Gaussian assumption is not appropriate. In this context, a performance metric adapted to non-parametric PDF should be used instead of CME (e.g., Wasserstein distance) to assess whether reconstruction performance improvements can still be achieved.”*

2. **The new approach was only compared to the benchmark approach such as the model democracy approach. However, various approaches that generate unequal weights have already been proposed. The authors also gave an overview of such kind of approaches. Why do not the authors compare the performance of their new approach to other unequally weighted approaches?**

We thank the reviewer for highlighting this point. We have clarified the manuscript on this.

Other established weighting approaches referenced in the introduction are used as benchmarks of our analysis (see Sec. 4.2.3). The CME-ClimWIP score derives from the ClimWIP approach (e.g., Knutti et al., 2017), by replacing a metric based on mean square error by the CME. Also, the CME-ClimWIP score can be interpreted as a Bayesian model averaging (BMA) approach, which relies on computation of the global model evidence. Specific values of the shape parameters (e.g., $\sigma_{D}=1$ and $\sigma_{S}=\infty$) yield the classic BMA expression. More details on this point are provided in Appendix A of the manuscript.

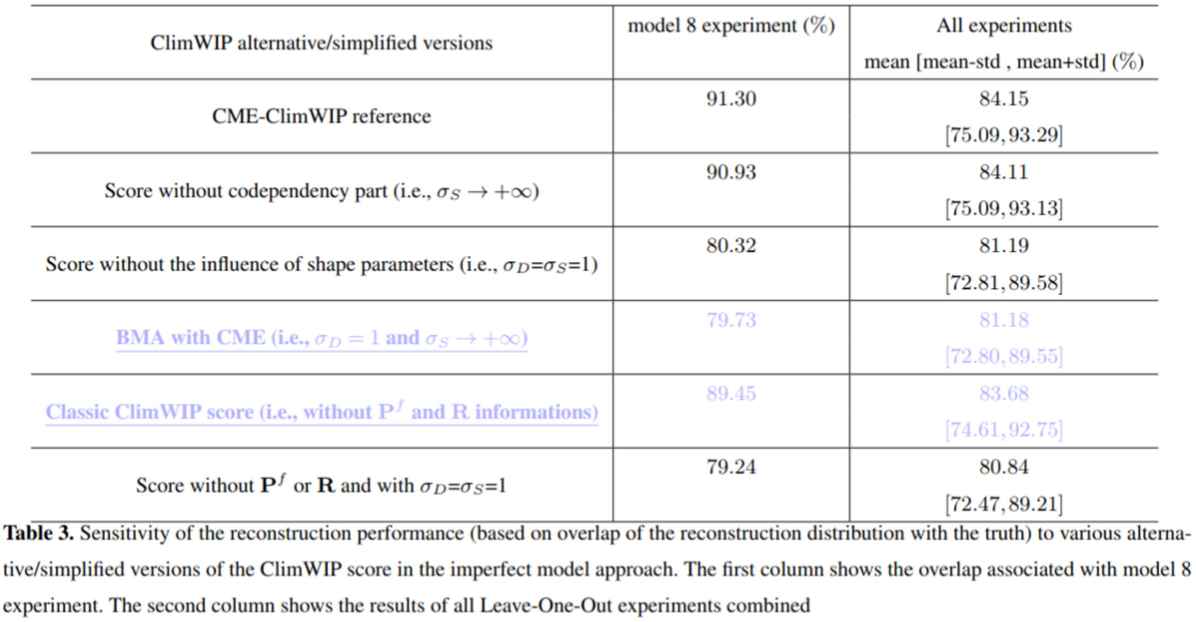
We have provided more details in the text to explicitly deal with these comparisons. In particular, Section 4.2.3 has been renamed and updated to provide further clarification. The text has been modified as follows:

Sec. 1, Page 3, Lines 63-65: *“The proposed methodology is relevant for various reasons. It effectively combines the advantages of multiple weighting methods mentioned earlier. Moreover, it is integrated within a data assimilation framework, enabling a thorough consideration of the described dynamics.”*

Sec. 2.2.1, Page 7, Line 182-185: *“Notably, higher values for these parameters bring the score closer to the model democracy while when $\sigma_{D}=1$ and σ_{S} tends to ∞, the BMA expression is retrieved. In particular, the numerator in Eq. (7d) corresponds to a formulation of model evidence typically calculated for model selection problems (see Appendix A for more details on this specific point). Classic ClimWIP and BMA approaches are special cases of the CME-ClimWIP score which are discussed in Section 4.2.3.”*

Sec. 4.2.3, Page 16, Lines 332-333, modified title section: “"Evaluating the reconstruction performance of alternative versions of the CME-ClimWIP score”.

Sec. 4.2.3, Page 16, Lines 338-341:*"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 whether or not the influence of each component (Table 3). Note here that two alternative versions represent weighting strategies mentioned in the introduction. Hence, BMA formalism (with model evidence estimated using successive CME) and the ClimWIP approach (not considering the covariance error matrix) are presented (refer to rows 5 and 6, respectively, in Table 3)."*



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]
Score without codependency part (i.e., $\sigma_S \to +\infty$)	90.93	84.11 [75.09, 93.13]
Score without the influence of shape parameters (i.e., σ_D=σ_S=1)	80.32	81.19 [72.81, 89.58]
BMA with CME (i.e., $\sigma_D = 1$ and $\sigma_S \to +\infty$)	79.73	81.18 [72.80, 89.55]
Classic ClimWIP score (i.e., without P^f and 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 reconstruction distribution with the truth) to various alternative/simplified versions of the ClimWIP score in the imperfect model approach. The first column shows the overlap associated with model 8 experiment. The second column shows the results of all Leave-One-Out experiments combined

3. **About Eq. 5, the authors demonstrate the contextual model evidence (CME) takes into account both reliability information and accuracy information. It is unclear which terms in Eq. 5 are associated with reliability or accuracy. Please give more description and discussion. Please also give discussion on why both reliability and accuracy should be considered in defining the CME. Notice that most existing data-driven model used a loss function which is solely a function of mean square error.**

We thank the reviewer for suggesting clarification of this point.

The CME is a function of mean square error which is normalized by the covariance matrix (Metref et al. 2019). In (5) of the manuscript, the mean square error term **(y_k - H(x^f_k))(y_k - H(x^f_k))^T** measures the accuracy of the forecast with respect to the observation value at time *k*. The covariance term **(H P^f_k H^T + R)^-1** provides information on the uncertainty (assumed Gaussian) of both the forecast state (the matrix **P^f_k**, derived from a DA cycle) and the observation (the matrix **R**) which reflects their reliability within the DA process. Hence, the error variance varies with time, which contrasts with a metric solely based on mean square error, where the error variance is assumed to be independent of the state of the system.

Recent studies (Metref et al., 2019; Ruiz et al., 2022) show that incorporating both prior information and uncertainty, the CME is more effective than using RMSE alone in selecting the correct model. In our study, tests were conducted using mean square error instead of CME. Results in Table 3 ("Score without P^f_k and R information") demonstrate lower reconstruction performance with this approach, highlighting the advantage of using CME in our framework (see lines 338-339 in the revised manuscript).

We clarify this point in the manuscript as follows:

***Sec. 2.1 - Page 5 - Lines 121-129**: "At time k, the CME is defined as a mean square error function between the forecast and the observation [represented by the term **(y_k - H(x^f_k)) (y_k - H(x^f_k))^T** in (5)], which is normalized by the covariance matrix uncounting for both forecast and observations error [denoted as the term **(H P^f_k H^T + R)^-1** in (5)]. It has been demonstrated that by considering both the prior and data uncertainty, the CME exhibits greater effectiveness in selecting the accurate model from a pool of models compared to using RMSE alone (Metref et al., 2019 and Ruiz et al., 2022)."*

4. For the data-driven approach, how long the historical data is sufficient for a robust estimation of the forward propagator?

In data-driven approach, the skill of the forecasts depends on the quality and representativity of the catalog (Lguensat et al., 2017). In our study, the optimal catalog size for obtaining reliable forecast estimates varies with each model version, depending on the complexity of its dynamics (e.g., the degree of nonlinearity). We find that a catalog size of 400.000 samples ensures robust forecasts for all model versions.

We add some clarifications in the Section 3.3 of the manuscript as follows.

***Sec. 3.3, Page 11, Lines 254-255:** "The methodology has been tested with different catalog sizes. A catalog size of 400,000 years ensures robust data-driven forecasts estimates in all the experiments."*

5. Section 2.2.2, what is the difference between the method used here and the widely used multi-linear regression?

In Section 2.2.2, we explain how the weights for each score are applied using new individual model simulations [see (9) below].

$$p(\widetilde{\omega}|\mathcal{M}^{(0)},\ldots,\mathcal{M}^{(L)},y) = \sum_{i=0}^{L} p(\widetilde{\omega}|\mathcal{M}^{(i)}) \cdot w_{\text{score}}^{(i)} \quad (9)$$

where $p(\widetilde{\omega}|\mathcal{M}^{(0)},\ldots,\mathcal{M}^{(L)},y)$ denotes the final reconstructed probability density function (PDF) of the normalized AMOC (noticed as $\widetilde{\omega}$ for the normalized version) for a specific score. This is obtained from the application of the model weight (denoted $w_{\text{score}}^{(i)}$) to the PDF of the simulations for model i (denoted $p(\widetilde{\omega}|\mathcal{M}^{(i)})$). Here, the reconstructed PDF differs from the multi-linear regression by considering the PDF of model simulations instead of continuous variables typically used as predictors and response variables in linear regression.

Section 2.2.2 has been completely revised as another reviewer requested further clarification on it. The updated version (including Figure 1) is as follows:

***Sec. 2.2.2, Pages 7-8, Lines 200-216 : "**Here, the methodology for one experiment consists in two successive stages (Fig. 1). The first step corresponds to a training stage where AnDA is applied to obtain the CME time series associated with each candidate model (left panel of*

Fig. 1). From them, CME-based scores are calculated following (7d-e-f) while the three benchmark scores are derived from(7a-b-c) (bottom center panel of Fig.1). The second step is a testing stage consisting in applying the six weighting approaches on new individual model simulations and comparing their effectiveness to suitably reconstruct the true one (right panel Fig. 1).

Considering probability density functions (PDFs) of model simulations, the reconstructed PDF is calculated as follows

$$p(\widetilde{\omega}|\mathcal{M}^{(0)},\ldots,\mathcal{M}^{(L)},y)=\sum_{i=0}^{L}p(\widetilde{\omega}|\mathcal{M}^{(i)})\cdot w_{\text{score}}^{(i)} \qquad (8)$$

where $p(\widetilde{\omega}|\mathcal{M}^{(0)},\ldots,\mathcal{M}^{(L)},y)$ *denotes the reconstructed PDF of the normalized AMOC (denoted as* $\widetilde{\omega}$*),* $p(\widetilde{\omega}|\mathcal{M}^{(i)})$ *represents the individual PDF associated with model i and* $w_{\text{score}}^{(i)}$ *is the weight of model i for a given score, depending on y.*

To evaluate the skill of the reconstructed PDF , its overlap (in terms of %) with the PDF of the truth is calculated [following similar formalism as the climatological score $w_{\text{climato}}^{(i)}$*, (7b)]. The resulting score performance for the six approaches can then be compared."*

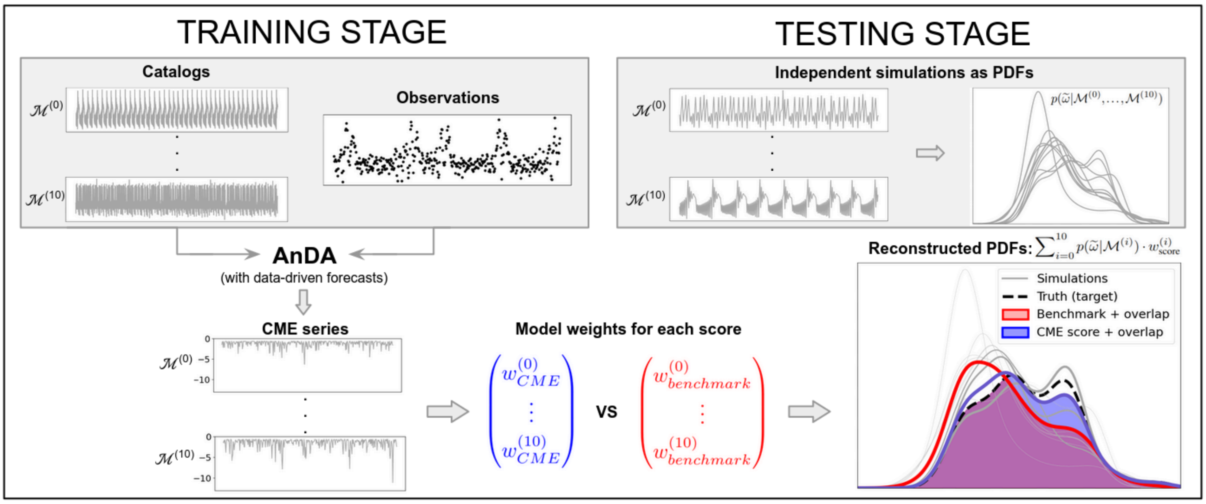

TRAINING STAGE
Catalogs
$\mathcal{M}^{(0)}$
$\mathcal{M}^{(10)}$
Observations
AnDA
(with data-driven forecasts)
CME series
0
−5
−10
Model weights for each score
$w_{CME}^{(0)}$
$w_{CME}^{(10)}$
VS
$w_{benchmark}^{(0)}$
$w_{benchmark}^{(10)}$
TESTING STAGE
Independent simulations as PDFs
$p(\widetilde{\omega}|\mathcal{M}^{(0)},\ldots,\mathcal{M}^{(10)})$
Reconstructed PDFs: $\sum_{i=0}^{10}p(\widetilde{\omega}|\mathcal{M}^{(i)})\cdot w_{\text{score}}^{(i)}$
Simulations
Truth (target)
Benchmark + overlap
CME score + overlap

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: model weights are calculated for two illustrative scores: one based on CME and another based on a benchmark method (blue and red, respectively). Right panel: individual PDF of new model simulations (grey) are weighted following (8) to obtain a reconstructed PDF associated with both competing scores (blue and red). The quality of both reconstructions is assessed by measuring the overlap (blue and red shadings) with the true PDF (dashed black line) which allows comparison of the performance of the two scores. Here, the CME score shows a higher overlap with the true distribution, highlighting a better reconstruction compared to the benchmark score.*

Grammatical errors:

1. **Ln25, "produces" -> "produced"**

Thank you. We corrected this in the revised version of the manuscript.

Reviewer 2

This manuscript describes the use of the Bayesian data assimilation framework to estimate the weights for ensemble of models.

Using the weights hence obtained, the authors compute the climatological distributions during independent experiments and compare its performances to other well-known weighting methods.

The manuscript is well-written, presents a novel approach and is in general well organized. One of the assets of this paper is that it proposes a methodology which is - to my knowledge - directly applicable to real world predictions (thanks to AnDA), however with the caveats presented in the conclusion.

Therefore I recommend accepting it for publication once the comments below are addressed.

We would like to thank the reviewer for the comments that helped to improve the clarity of the manuscript.

1) The authors emphasize that the CME method is based on the models "short-term" dynamics, i.e. that the CME weights are obtained during the DA cycles. This is also emphasized in the Appendix A where it is said that this approach provides "more informative insights into current conditions, including forecast states with their uncertainties". However, in the end, it is the average of the CME over a number of K of DA cycles which is used to compute the weights, covering the "attractor" of the models. So I don't understand this claim. To me a "climatology" of the CME is constructed and used and so I have some trouble understanding the difference on this ground with climatology-based weighting. I would not call this weighting method a local one.

We appreciate the reviewer for pointing out these confusing ideas in the manuscript.

First, we focus on how weights are constructed using local skills and highlight their benefits. Then, we clarify how our method is different from the traditional climatology-based approach, which has been the confusing point.

The CME captures the local forecast skill, i.e., dependent of the current model condition. Then, the weights are defined by a scalar based on a (long) statistic of local CME values. This procedure has some advantages. As pointed out in the Appendix, the sum of local CME values over the observed period [see CME-ClimWIP formalism in(7d)] is a robust estimation of the model evidence defined as the global log-likelihood of the observations (Carrassi et al., 2017). This metric is usually calculated for model selection purposes (Reich and Cotter, 2015). However, estimating it in practical applications is complex, particularly with nonlinear geophysical models. Sequential DA addresses this challenge through the iteration of successive log-likelihood information using the CME. This involves assuming a Gaussian distribution locally over time, which is in general more appropriate than assuming it on global climatology.

Without the assimilation stage, the evaluation cannot be performed using the local properties since the current condition of the system is not accessible. In this case, a rougher evaluation such as a comparison between climatologies (which are not always Gaussian) can be done [see the climatology-based score described in (7b)].

However, we agree with the reviewer that the proposed method cannot be strictly defined as a “weighting local” one, since it does not vary over time. Instead, the method is qualified as a “weighting method taking into account local dynamics". This has been clarified in the manuscript.

A methodological perspective of this work would be to further take advantage of the locality of the CME by calculating time-varying weight.

We have clarified that in the manuscript.

Sec. 1, Page 3, Lines 69-72: *“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 a metric of model performance. The information of the local likelihood is then summarized following different score formalisms to yield weights. Here also, the scores take into account the model dependency as in the formalism of ClimWIP (Knutti et al. 2017). The weights are tested by applying them to long-term simulations.”*

Sec. 2.1 - Page 3 - Lines 85-86: *“Here, we evaluate the model performance based on its short-term dynamics. For this purpose, initialized model forecasts are crucial to estimate the accuracy of the dynamics with regards to observations. ”*

Sec. 2.2.1, Page 6, Lines 160-161: *“The climatological score is based on the comparison of model and observation distributions. Here, there is no assimilation process, which means that this score does not evaluate the local dynamics.”.*

Sec. 4.1.2, Page 14, Line 292-294: *“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). These results emphasize that identifying the correct model is more effective using local dynamics (thanks to DA framework) rather than a climatological-statistics.”.*

Sec. 4.2.2, Page 15, Lines 319-321: *“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). [...] These results highlight the ability of using local-dynamics over climatology-statistics for accurately differentiating between models for our reconstruction problem.”*

Sec. 5, Page 18, Lines 374-375: *“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. This emphasizes the valuable information contained within the short-term dynamics, rather than the general information provided by climatological statistics, enabling efficient differentiation between models.”*

Sec. 5, Page 18, Lines 388-390: *In the current study, the weights are fixed and do not change over time, as long-trem average of stationary series are reconstructed. From a methodological standpoint, there is potential to better leverage the time-local properties of the CME by computing time-dependent weights. This would be especially relevant in a non-autonomous framework (e.g., in the context of climate changes).*

2) "Section 2.2.2: Since this is the heart of the methodology, I would have expected more details about it. In particular, the link between Eq. (7d) (and the equations below) and Figure 1 should be made."

We agree with the reviewer that Section 2.2.2 needs more details. We propose to rewrite the full section by adding more methodological details. Also, we modify Figure 1 by specifying in the right panel the formula used to obtain the final PDF and bringing more details in the caption.

Sec. 2.2.2, Pages 7-8, Lines 200-216: *"Here, the methodology for one experiment consists in two successive stages (Fig. 1). The first step corresponds to a training stage where AnDA is applied to obtain the CME time series associated with each candidate model (left panel of Fig. 1). From them, CME-based scores are calculated following (7d-e-f) while the three benchmark scores are derived from(7a-b-c) (bottom center panel of Fig.1). The second step is a testing stage consisting in applying the six weighting approaches on new individual model simulations and comparing their effectiveness to suitably reconstruct the true one (right panel Fig. 1).*

Considering probability density functions (PDFs) of model simulations, the reconstructed PDF is calculated as follows

$$p(\widetilde{\omega}|\mathcal{M}^{(0)},\ldots,\mathcal{M}^{(L)},y)=\sum_{i=0}^{L}p(\widetilde{\omega}|\mathcal{M}^{(i)})\cdot w_{\text{score}}^{(i)} \qquad (9)$$

where $p(\widetilde{\omega}|\mathcal{M}^{(0)},\ldots,\mathcal{M}^{(L)},y)$ *denotes the reconstructed PDF of the normalized AMOC (denoted as* $\widetilde{\omega}$ *),* $p(\widetilde{\omega}|\mathcal{M}^{(i)})$ *represents the individual PDF associated with model i and* $w_{\text{score}}^{(i)}$ *is the weight of model i for a given score, depending on y.*

To evaluate the skill of the reconstructed PDF , its overlap (in terms of %) with the PDF of the truth is calculated [following similar formalism as the climatological score $w_{\text{climato}}^{(i)}$*, (7b)]. The resulting score performance for the six approaches can then be compared."*

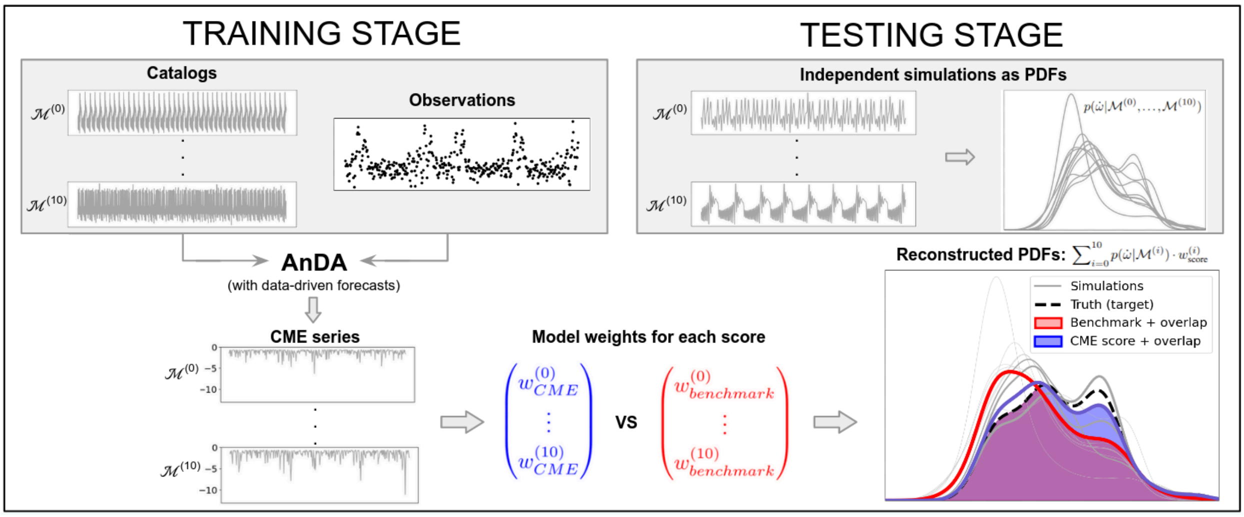


***Figure 1.** Schematic of the two-stages methodology for one experiment using 11 models. Left panel: training stage performing AnDA with the 11 catalogs and a single set of observations leading to CME computation. Bottom center: model weights are calculated for two illustrative scores: one based on CME and another based on a benchmark method (blue and red, respectively). Right panel: individual PDF of new model simulations (grey) are weighted following Eq. 9 to obtain a reconstructed PDF associated with both competing scores (blue and red). The quality of both reconstructions is assessed by measuring the overlap (blue and red shadings) with the true PDF (dashed black line) which allows comparison of the performance of the two scores. Here, the CME score shows a higher overlap with the true distribution, highlighting a better reconstruction compared to the benchmark score.*

3) Line 225: "(which avoids computational divergence for certain attractors with highly nonlinear dynamics...)" I do not understand this. Does that mean that some models are unstable? Or that their nonlinearity is too high compared to the others? Or something else? Please clarify.

We agree with the reviewer that the sentence was unclear.

Models exhibit different levels of nonlinearities and different occurrences of extreme values. A large number of analogs is sometimes necessary, in particular near extreme values. This is crucial for maintaining the stability of the forward operator, which relies on the estimation of coefficients for local-linear regression (LLR). In our study, the number of analogs is set to a fixed and high value (10,000 analogs). This ensures that forecast distributions are suitably estimated for the eleven models.

We have clarified this aspect in the manuscript as follows:

***Sec. 3.3, Page 11, Lines 255-258:** "For each of the 200 EnKF-members of an AnDA cycle, forecast distributions are estimated using a fixed number of 10,000 analogs. These allow stability of the forward propagator algorithm especially near unfrequently visited regions, often associated to extreme values, where tracking suitable analogs can be more challenging (Lguensat et. al, 2017).*

References cited in the reviews (in blue: the new ones)

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.

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.

Jarque, Carlos M., and Anil K. Bera. "Efficient tests for normality, homoscedasticity and serial independence of regression residuals." Economics Letters 6.3 (1980): 255-259.

Knutti, Reto, Jan Sedlácek, BenjaminM 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

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.

Reich, S. and Cotter, C.: Probabilistic forecasting and Bayesian data assimilation, Cambridge University Press, https://doi.org/10.1017/CBO9781107706804, 2015

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

Van Leeuwen, P. J., 2009: Particle filtering in geophysical systems. Mon. Wea. Rev., 137, 4089–4114, doi:10.1175/2009MWR2835.1.