The referee's comments below are in italics, our answer in plain font in blue

This manuscript presents a new paleoclimate reanalysis. An offline paleoclimate data assimilation is first performed by updating the priors from ensemble simulations of NorCPM with proxy recodes. The offline three-dimensional atmospheric zonal and meridional wind analyses are then incorporated into the NorCPM through nudging. The wind-nudged ensemble simualtions captures ocean variability and exhibits pronounced multi-decadal variability. These results highlight the role of wind-driven ocean adjustment in shaping low-frequency climate variability.The manuscript is well written. The generated fully copled ocean-atmosphere paleoclimate reanalysis over the past centureis could be a valuable contribution to the climate community. I have several specific comments as below.

We thank the reviewer for their positive overall assessment of our manuscript and for their constructive comments, which we address point by point below.

I'd like suggest the authors use the generally adopted terminology of data assimilation. Bold lower-case x is used for state vector, italic and lower-case h for PSM, and bold and upper-case H for the Jacobian of PSM, etc. Please note that h in innovation and H in Kalman gain should be different.

We thank the reviewer for this suggestion and try to follow the Ide et al 2017 terminology. Please note that while we use **x** as state vector, we use **X**a (bold non-italic upper-case) for the matrix (ensemble of state vectors, as it is the common terminology in mathematics. The ensemble mean (overbar **x**a) is lower case bold non italic (vector) . **y** for the observation vector, italic *h* for the observation operator, i.e. the proxy system model, and bold upper-case **H** for its linearisation (Jacobian). In particular, the innovation will be written as **y** − *h*(**x**^b), while **H** will be used in the Kalman gain, **K** = **P**^b**H**^T(**HP**^b**H**^T + **R**)^−1.

Lines 120-130, 1) how the inflation value, 'increased by a factor of two', is determined? 2) Please give more details of the observation error inflation. When the observation error inflation is applied, and how large the inflation value is?

We thank the reviewer for this comment. There are many flavours of inflation, for instance, the standard multiplicative inflation (Anderson and Anderson 1999), adaptive inflation (Anderson 2007) and relaxation to the prior (Whitaker et al. 2012), EnKF-N (Bocquet et al. 2015). The two inflation-like methods used here are derived from (Sakov et al., 2012). In the square root filter version of the Ensemble Kalman Filter, the analysis works in two steps: the update of the ensemble mean and the update of the ensemble anomaly. The first method we use (called r-factor) inflates the observational error when updating the ensemble anomalies (Eq. 4), while the ensemble mean is updated with the original observational error. This ensures that the update of the ensemble mean fully exploits the observation, but prevents a spurious collapse of the ensemble anomaly (e.g. due to sampling error). This method works well with a sparse and inhomogeneous observation network unlike multiplicative inflation (Anderson 2001). The value of two was not tuned specifically for this study. It is the standard setting advocated by Sakov et al. 2012 and used in all NorCPM simulation, (Counillon et al., 2016; Bethke et al., 2021; Wang et al., 2025).

The second inflation-like approach is referred to as the pre-screening of observations. This procedure, also inherited from the TOPAZ4 system (Sakov et al., 2012), aims to keep the update consistent with the model physics, as represented by the ensemble spread of the prior (i.e., the forecast error). Large innovations can arise for different reasons: an underestimation of the observational error, for instance related to the proxy system model or to dating uncertainties, or model biases that make the observation incompatible with the model state. In both cases, assimilating the observation with its original error would produce an increment far outside the range of the prior ensemble, which may lead to unphysical updates and assimilation shocks with online assimilation. Rather than discarding such observations, we reduce their impact by increasing their observational error as follows:

$$\theta^{\widetilde{2}}_{obs} = \sqrt{\left(\theta^{2}_{ens} + \theta^{2}_{obs}\right)^{2} + \left(\frac{1}{Kfactor}\theta_{ens}d\right)^{2}} - \theta^{2}_{ens}$$

Where $\theta^{\widetilde{2}}_{obs}$ corresponds to the modified observation error, from the original observation error θ^{2}_{obs}. $\theta^{2}_{ens}$is the standard deviation of the ensemble (i.e., the forecast error). d corresponds to innovation. In such a way, the analysis remains within k*times the ensemble spread (Kfactor). In NorCPM, $Kfactor$ it is set to two the value recommended by Sakov et al. 2012.

The two moderation techniques will be further detailed in the revised manuscript.

Lines 140-145, 1) are the same ensemble priors used for each DA step, or the 30 ensemble priors are used as ensemble priors? 2) The two references used simplified models. But as shown by Lei et al. (2026), the information from analyses can be propagated along with online cycling data assimilation.

For each assimilation step (i.e., each year), the prior consists of the 30 members of the transient NorESM1 ensemble for that specific year, covering April to March at monthly resolution. The prior therefore, differs from one year to the next and includes the externally forced signal specific to that year, while the 30 members sample the internal variability. In other words, the same 30 transient simulations are used throughout the reconstruction, but the prior at each step is taken from the corresponding year of these simulations. As the approach is offline, the prior is not updated with the analysis of the previous step. This will be clarified in the revised manuscript.

As for the online assimilation, we thank the reviewer for this comment and for pointing us to the study of Lei et al. (2026). In the revised manuscript, we will rephrase this statement to add references to more recent studies using more complex models (Perkins and Hakim, 2017; Lei et al., 2026).

Lines 152-153, how is the flow-dependent background error covariances constructed from the 30-member ensemble different from that constructed from the "analog" ensembles (Sun et al. 2022)? Please give some intuitive expectations.

We thank the reviewer for this comment. There are similarities between the two approaches. In our framework, the prior for a given year consists of the 30 members of the transient

NorESM1 ensemble for that year. All members share the same external forcing, so the ensemble mean contains the externally forced signal specific to that year, and the year-to-year changes in the dynamical covariance matrix reflect the influence of the external forcing on the variability, for instance, after major volcanic eruptions. They do not, however, track the actual state of internal variability. In contrast, the analog offline EnKF of Sun et al. (2022) selects, at each assimilation time, the model states of a long simulation that are closest to the proxy observations. The prior is therefore conditioned on the observed state, which provides both a more accurate prior mean for internal variability and covariances representative of the actual dynamical regime. The analog selection can thus be seen as a refinement of the prior before the EnKF update. Intuitively, we would therefore expect the analog approach to better constrain internal variability, while our approach better represents the forced component and the full spread of internal variability. This difference should be most important when the observational network is very sparse, for instance during the earliest centuries of the reconstruction. In such cases, the analog selection relies on only a few observations, so that the selected prior, particularly far from the available records, tends towards the climatology of the control simulation and does not contain the response to the external forcing. In our framework, the prior still provides the forced response of the model, for example after major volcanic eruptions. We will clarify this distinction in the revised manuscript and discuss the analog approach of Sun et al. (2022). Particularly, applying the analog method on the static ensemble seems to us a promising perspective.

Lines 160-165, this procedure is similar to the hybrid gain proposed by Sun et al. (2024). Do you have an explicit sample-estimated P_d? How the hybrid background error covariance matrix P is applied to the solution of ensemble analyses, as shown by equations (3) and (4)?

We thank the reviewer for this comment. Our approach is similar in spirit to the hybrid gain approach of Sun et al. (2024), as both combine flow-dependent covariances estimated from a limited ensemble with static covariances estimated from a long simulation. However, the two formulations are different (while they should converge to a similar solution when the observation error dominates). In the hybrid gain approach of Sun et al. (2024), two Kalman gains are computed separately from the flow-dependent and static covariances and then linearly combined.

We use the “classic” hybrid background error covariance, in which the covariances are combined first, $\mathbf{P} = \beta\mathbf{P}^d + (1-\beta)\mathbf{P}^s$, and a single Kalman gain is computed from $\mathbf{P}$ (Eq. 2). This approach was introduced by Hamill and Snyder 2000 and widely used in numerical weather prediction and operational oceanography. We use the ensemble formulation proposed in Counillon et al. (2009), which is mathematically equivalent.

As such $\mathbf{P}^d$ and $\mathbf{P}^s$ are not formed explicitly. In practice, the hybrid covariance is obtained by concatenating the ensemble anomalies of the two ensembles, scaled by $\sqrt{\beta}/\sqrt{(N-1)}$ and $\sqrt{(1-\beta)}/\sqrt{(M-1)}$ respectively, into an augmented anomaly matrix. N corresponds to the length of the dynamical ensemble (i.e., 30) while M is the number of years of the simulation used for the static covariance (i.e., 300). To note, the static anomalies only contribute to the estimation of the Kalman gain; they are not updated and do not form part of the posterior ensemble. These details, together with the relationship to the hybrid gain approach of Sun et al. (2024), will be added to the revised manuscript.

Lines 205-206, do you expect wider localization lengthscales for P_s than P_d?
Table 2, what's the unit for 'Min. length'?

We thank the reviewer for this question. We would indeed expect a wider localization for **P**^s than for **P**^d, as the static covariances are estimated from a much larger sample (300 years versus 30 members) and are therefore less affected by sampling errors if we were doing the update independently. However,hybrid covariance and localisation address the same issue (sampling error). Therefore, we rather formulate the hybrid covariance that stabilises (the smaller in ensemble size) flow ensemble covariance with the (the larger in ensemble size) static and fine-tune a posteriori the localisation to maximise the usage of the observation. Ménétrier and Auligné (2015) and Ménétrier et al. (2015) formulated a way to perform hybridisation as a linear filtering problem of the background error covariance matrix to optimise both the localisation and the hybridisation coefficients simultaneously.

As for the "Min. length", it corresponds to the minimum number of data points when calibrating the proxy system models. To be more explicit, we will clarify this, and change for "minimum number of data points". When the criterion is not respected, the record is not included in the system.

Lines 382-385, observation errors are important for paleoclimate data assimilation. How's the sample size for each proxy record that is used for PSM calibration. Are there enough samples for this observation error estimation? Are the samples assumed to be independent? If so, is the assumption valid for the same proxy record?

We thank the reviewer for this important comment. The sample size for the calibration of each proxy record corresponds to the overlap between the record and the calibration dataset, which starts in 1950, 1958 or 1979 depending on the proxy type and region (Table 2). The calibration period therefore covers at most about 50 years. To ensure a minimum robustness of the calibration, we require at least 21 data points for all proxy types, except for snow-accumulation records, for which the minimum is relaxed to nine data points following Medley et al. (2018a), as these records are directly compared with precipitation. We agree that such sample sizes are limited for estimating the observation error, and therefore we might underestimate the observation error. This is precisely why we do not estimate it from the residuals over the full calibration period. Instead, we use a Monte Carlo cross-validation approach. For each record, 1000 realizations are generated by randomly selecting 80 % of the available data to calibrate the regression and using the remaining 20 % to compute the residual error. The final observation error is the mean error variance over the 1000 realizations. This provides an out-of-sample estimate that accounts for the uncertainty associated with the limited calibration data, and leads to a more conservative weighting of the proxies in the assimilation. As stated in the original manuscript, the resulting increase in the observation error is about 20 % (in standard deviation) compared with the residuals over the full calibration period. We consider it as moderate and consistent with the expected effect of finite sample sizes in a linear regression with one or two predictors fitted on 20-50 data points. A much larger increase would indicate that the calibrations overfit the data. This suggests that the calibration periods are sufficiently long to provide a reasonable estimate of the observation error. In addition, only records whose regression slope is statistically significant at the 95 % level are assimilated, which excludes records whose relationship with climate cannot be robustly established given their sample size.

Regarding independence, the 1,000 calibration subsets are indeed not independent from each other, as they are drawn from the same data points and largely overlap. This corresponds to a repeated random-subsampling cross-validation. The number of realizations only reduces the sensitivity to any particular data split. It does not add independent information, which remains limited by the number of data points of each record. Since only the mean over the realizations is used, and not their spread, the dependence between subsets does not affect the estimated observation error.

Lines 400-405, the proxy records with large innovations could contain valuable information, especially for the extreme events that is very difficult to simulate by numerical models. For offline paleoclimate data assimilation, the imbalance or instability is not associated with model simulation, so that is it possible to include such extreme-value proxy records?

We thank the reviewer for this comment. We agree that extreme proxy values may contain valuable information on extreme events, which climate models have difficulties to simulate, and that, in an offline framework, a large increment in one year cannot propagate to the following years. However, throughout the development of the reanalysis, we paid particular attention to keeping the reconstruction consistent with the model physics, and therefore deliberately adopted a conservative approach. As each year is reconstructed independently, a single extreme value can produce a large increment in one year, resulting in an abrupt and unphysical jump in the reconstruction. This is particularly important here, as the reconstructed winds are then imposed on the coupled model through nudging. Moreover, such values might reflect non-climatic factors. As noted in the original manuscript, a similar screening is used in previous reanalyses (e.g., Dalaiden et al., 2023; Valler et al., 2024).

By definition, values beyond 3.5 standard deviations are extremely rare, so this screening does not affect the ability of the reanalysis to reproduce the large-scale climate variability. Moreover, all other observations, including those associated with large innovations, are assimilated, with their impact moderated through the kfactor described above. Their information is therefore retained rather than discarded, while avoiding updates far outside the range of the model physics. We acknowledge, however, that it may limit the reconstruction of some extreme events, and we will mention this limitation in the revised manuscript, together with a clarification of the wording.

Lines 444-445, are the temperature and SST updated by the proxy? If yes, how abonut nudge to 3-D wind and mass fields?

We thank the reviewer for this question. Yes, temperature (including at upper levels) and SST are part of the state vector and are updated by the proxies in paleo-NorCPM1 offline. Constraining only the winds in the coupled simulations is a deliberate choice. The main motivation of the wind-nudging framework is to let the ocean respond dynamically to the reconstructed circulation while preserving its ability to feed back on the atmosphere through thermodynamics. Nudging temperature, or the mass field (e.g. surface pressure), in addition to the winds would strongly constrain the surface heat and freshwater fluxes, bringing the configuration close to an ocean model forced by prescribed atmospheric conditions, as in Dalaiden et al. (2025), and removing the coupled feedbacks we aim to represent.

Constraining SST would be even more restrictive, since it would directly prescribe the ocean surface state.

Lines 454-456, how about the nudging coefficient in equation (10)? Would a trapezoidal nudging coefficient help with synchronization?

We thank the reviewer for this suggestion. In our configuration, the nudging coefficient is constant in time, with a restoring timescale of 6 h. From our understanding, a trapezoidal temporal weighting (e.g., Lei and Hacker, 2015) makes the nudging stronger around the times at which the target is available, and is particularly suited when targets are only available at discrete times. With monthly targets, it could help smooth the transitions between consecutive months. However, as described in the manuscript, we tested several restoring timescales (one month, 10 days and one week) when nudging directly toward the monthly winds, which all led to weak synchronisation and a growing lag. Therefore, we do not expect that it would solve the main issue we encountered, namely the absence of sub-monthly information in the reconstructed winds.

Equation (13), is xi here the ensemble mean?

Yes, so we will correct to be consistent.

Figure 3, between 1500 and 1600, paleo-NorCPM1 and paleo-NorCRM1 wind nudging are very similar to each other. Does this indicate that the spin-up period for online data assimilation is around 100 years?

We thank the reviewer for this comment. The wind-nudged simulations only cover the period 1600–2000 and are therefore not available between 1500 and 1600. We realise that this was not clear in the figure. We will clarify this in the caption of Fig. 3 and in the text.

Figure 5, do you have explanations for the discrepancies between paleo-NorCMP1 wind nudging and LMR online before 1800?

We thank the reviewer for this question. Attributing these discrepancies is difficult, but we argue that they mainly reflect the fundamentally different methods used to represent the ocean, which become more important before 1800, when the proxy network is sparser compared with the more recent periods and none of the assimilated records directly constrains the ocean heat content. In LMR Online, the ocean memory is represented by a linear inverse model trained on last-millennium simulations, whereas in paleo-NorCPM1, the ocean heat content results from the dynamical response of a fully coupled model to the reconstructed atmospheric circulation via wind nudging. Since models are known to underestimate low-frequency variability compared with proxy-based estimates (e.g., Laepple et al., 2023; Hébert et al., 2025), the linear inverse model is expected to inherit this limitation, which may explain the weaker multi-decadal variability in LMR Online. The two products also rely on different climate models, and therefore on different ocean dynamics.