We thank the reviewer for the insightful comments. Our point-to-point responses are placed below each major comment and highlighted in blue.

General comments:

This manuscript presents a new prediction system based on NOAA GFDL's SPEAR physical model coupled to COBALTv3 ocean biogeochemical model. The authors evaluate historical reconstructions and initialized retrospective predictions of air-sea CO2 flux over 1984–2023. They report that initialization improves global prediction skill of air-sea CO2 flux for approximately two years, primarily through ENSO-related variability, and identify a multidecadal predictable component associated with Eastern Boundary Current upwelling and Southern Ocean mixed layer depth and sea surface temperature.

The development of a coupled prediction capability enabling carbon cycle variations is valuable, and the comparison of predictive skill against observation-based reconstructions and the model's own reconstruction provides further implications of predictive skill triggered by reference data uncertainty. This study requires a lot of work on setting up the prediction system, running ensemble of simulations, and conducting analyses. This manuscript is well designed and clearly written. A few aspects listed below need to be further clarified.

I wonder why the predictive skill of air-sea CO2 fluxes presented here is in general shorter than the previous studies (Lovenduski et al. 2019; Li et al. 2019; Ilyina et al. 2021…) revealed. The longer-predictive skill (>3 years) was attributed to non-thermal processes such as the circulation and biological processes, therefore, initialization of the ocean interior is critical for maintaining higher skill. The authors explained why they use the current assimilation method practically, but do the predictions have the potential to improve by adding more ocean interior data?

We argue that the predictive skill presented here is generally comparable to that of Lovenduski et al. (2019) and a range of studies summarized in Ilyina et al. (2021), though potentially lower than in Li et al. (2019). Note that predictive skill is evaluated by comparing forecasts against observation-based data products, whereas predictability is evaluated by comparing against the model's own reconstruction. These definitions are applied consistently across all referenced studies. In Lovenduski et al. (2019), gridpoint-wise predictive skill at a 1-year lead time is presented in their Figure 7; we present similar results in our Figure 2c. Although time coverage and reference data products differ between the two studies, the gridpoint-wise skill in our study is overall higher than in Lovenduski et al. (2019). Table 1 of Lovenduski et al. (2019) lists biome-averaged predictability (prediction) horizons in the rightmost column, ranging from zero to three years across all biomes (in parentheses). This is consistent with, if not shorter than, the horizons presented in our Figure 4b. Regarding predictability, Lovenduski et al. (2019) highlight regions with long predictability horizons of 9–10 years, such as the Southern Ocean and Atlantic high-latitudes. Our findings align with this, with the exception that our initialized predictions were run for 5 years, capping our derived predictability horizon at 5 years.

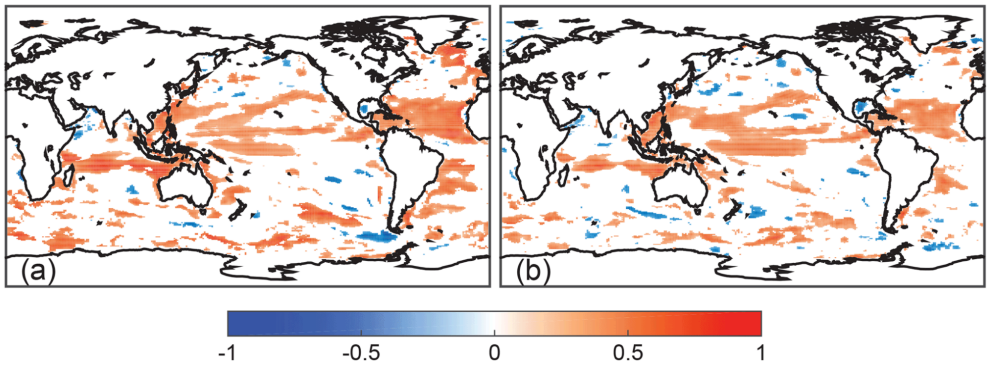


Figure 7 in Lovenduski et al. (2019) showing prediction skill at a lead time of 1 year based on ACC against SOM-FFN.

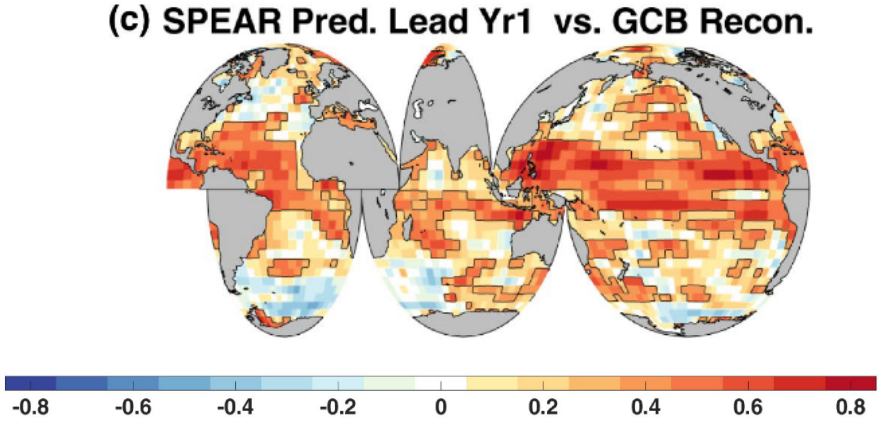


Figure 2c (this study) showing prediction skill at a lead time of 1 year based on ACC against the mean of the eight data products included in GCB2024.

In contrast, Figure 2A of Li et al. (2019) displays large regions with significant predictive skill at a 2-year lead time. However, it is unclear whether their correlations were computed from raw anomalies or detrended anomalies relative to SOM-FFN. We conducted parallel analyses using both approaches, showing that the predictive skill in our study is comparable (based on raw anomalies) or lower (based on detrended anomalies) relative to the values reported in Figure 2A of Li et al. (2019). Globally, Li et al. (2019) report an ACC of 0.71 (detrended ACC of 0.81) at a 2-year lead time (their Figure 1A), whereas our study yields 0.94 and 0.57, respectively (Figure 1f). This indicates that our system better predicts the observed long-term trend, while Li et al. better captures observed interannual variability. Notably, our detrended ACC matches the ensemble average of eight previous studies summarized in Ilyina et al. (2021) (0.57 vs. 0.58).

Nonetheless, we fully agree with the reviewer that initializing the ocean interior is critical for sustaining long-term skill. As discussed in Section 4.1, ocean interior assimilation in our current framework introduced spurious vertical velocities, particularly near the equator. Assimilating subsurface ocean data generated large T/S increments along the equator at depths of 70–150 m as well as at the surface, inducing unrealistic vertical motion. Biased trade winds further exacerbated this upwelling, leading to excessive surface nutrient delivery and primary productivity. Fortunately, ongoing work with

GFDL's Earth System Model (ESM4.1) using the Parallel Data Assimilation Framework (PDAF) demonstrates that the ESM4.1-PDAF system successfully assimilates ocean interior data with substantially reduced equatorial biogeochemical biases (pers. comm., Park and Jeon). In ESM4.1-PDAF reanalyses, peak equatorial T/S increments occur at much shallower depths (30–50 m), preventing localized vertical velocities from penetrating the nutricline. This encouraging result indicates that higher-resolution ocean models and refined DA methods can mitigate these issues. We are actively collaborating with the ESM4.1-PDAF team to incorporate subsurface initialization into our next-generation ESM-based prediction system.

The authors claimed in the abstract and conclusion of separating natural variability from anthropogenic variability for predictions. It is unclear how the separation was done. Besides CO2 forcing, aerosol forcing was also introduced, a simple linear trend might not lead to a clear separation of anthropogenic effects.

The direct quantification of predictive skill is in general clear, but the Average Predictability Time (APT) analysis and how they relate to different climate mode needs further clarification. Somehow it is mixed of natural variability and anthropogenic variability. It is not straightforward to understand why the APT decomposition is based on original time series and the correlations are calculated with detrended time series.

We thank the reviewer for these insightful comments, which address interconnected aspects of our analysis. Here, we clarify our approach using an extended Average Predictability Time (APT) framework. The principal strength of APT analysis is that it decomposes multivariate data into orthogonal components, each maximizing a distinct portion of total predictability. By design, the leading component represents the most predictable signal. When applied to raw time series, APT isolates a component dominated by externally forced variability (driven by CO_2 and aerosol forcing) that remains consistent across ensemble members and lead times. This represents the primary predictable signal identified by APT, exhibiting the longest predictability horizon (theoretically reaching the maximum forecast length in the absence of ensemble noise). However, forced variability does not represent skill gained specifically from initialization, as evidenced by Figure 6c, where ACC values are comparably high across all simulation types. Because APT components are orthogonal, sequentially removing leading components leaves subsequent components unaffected, enabling a clean separation between externally forced signals (the first component) and internally generated natural variability (subsequent components). During manuscript preparation, we also evaluated APT using detrended time series. APT reliably extracted the remaining components from the detrended dataset (i.e., new Component 1 aligned with original Component 2, new Component 2 aligned with original Component 3, etc.). However, standard linear detrending failed to fully remove nonlinear forced signals (such as volcanic impacts), causing residual forced variability to contaminate the first component. Conversely, polynomial detrending introduced spurious multidecadal artifacts. Consequently, performing APT on the raw time series proved to be the most robust approach for isolating forced predictability from initialized natural predictability. We will expand our explanation of these methodological choices in the revised text and include the detrended APT analysis in the Supplemental Material.

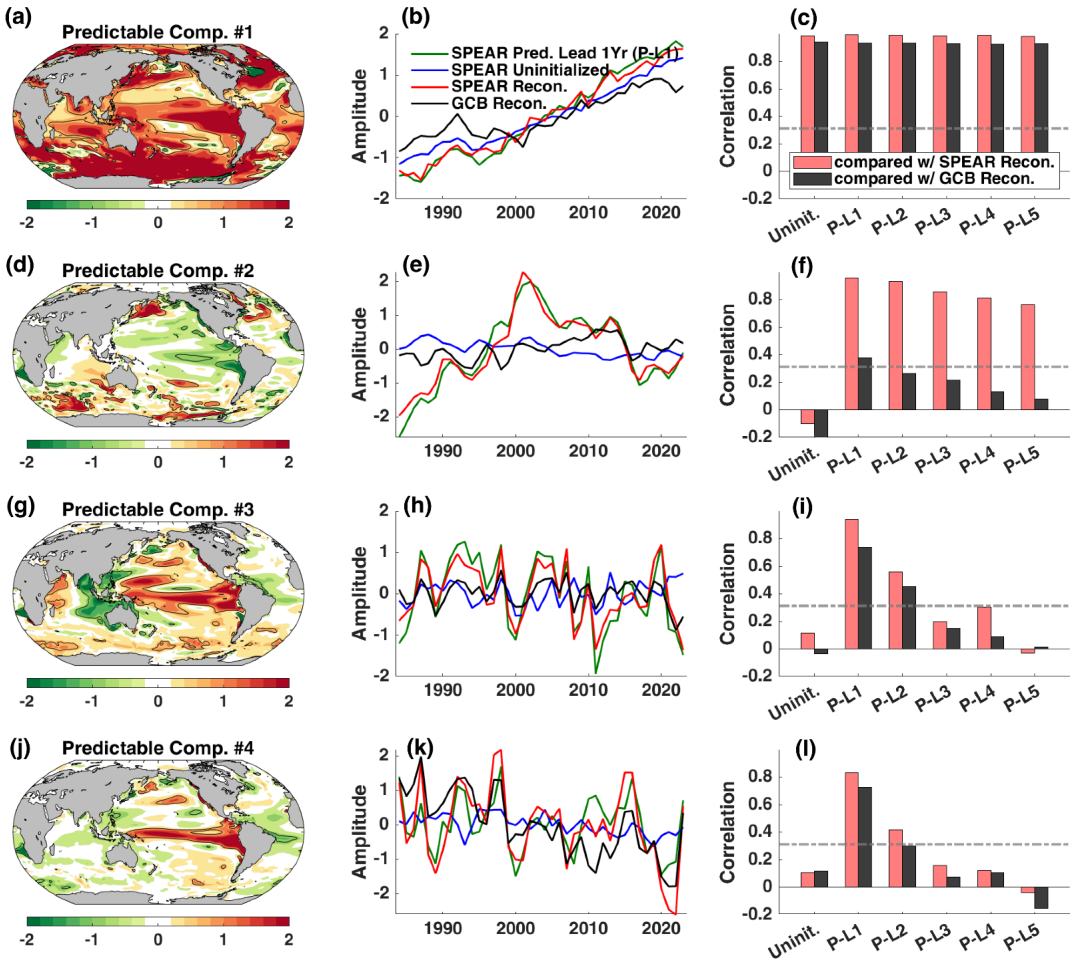
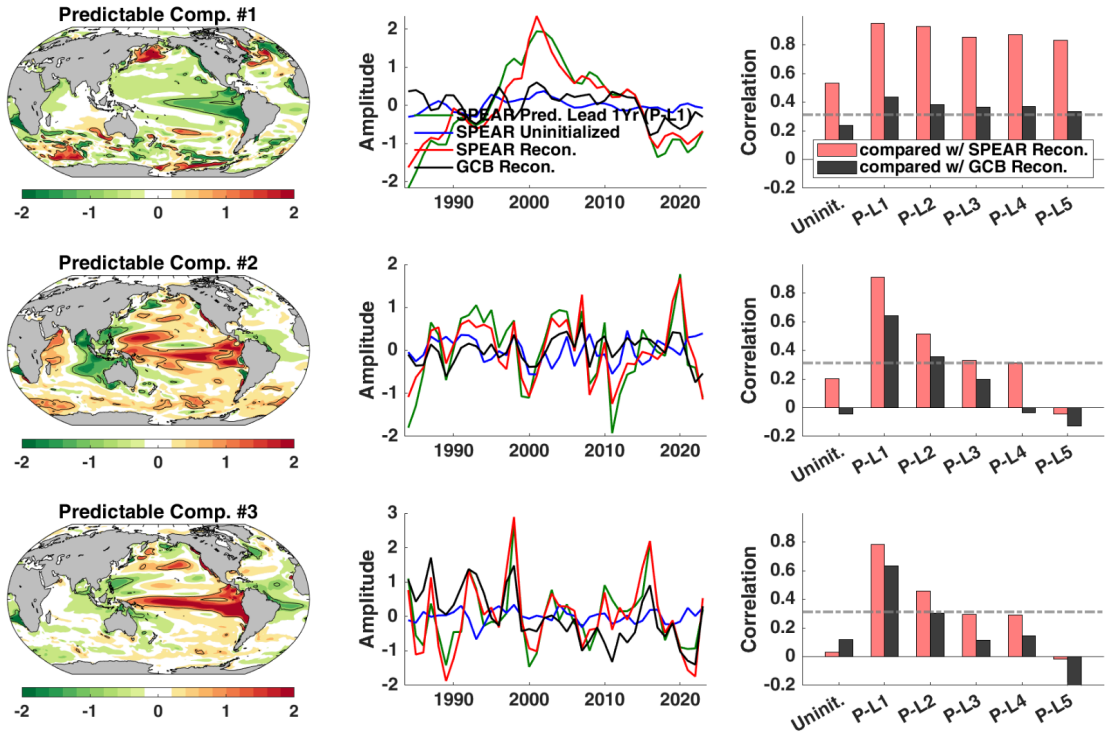


Figure 6 (this study) showing the four leading predictable components from APT analysis using the original or raw anomalies.

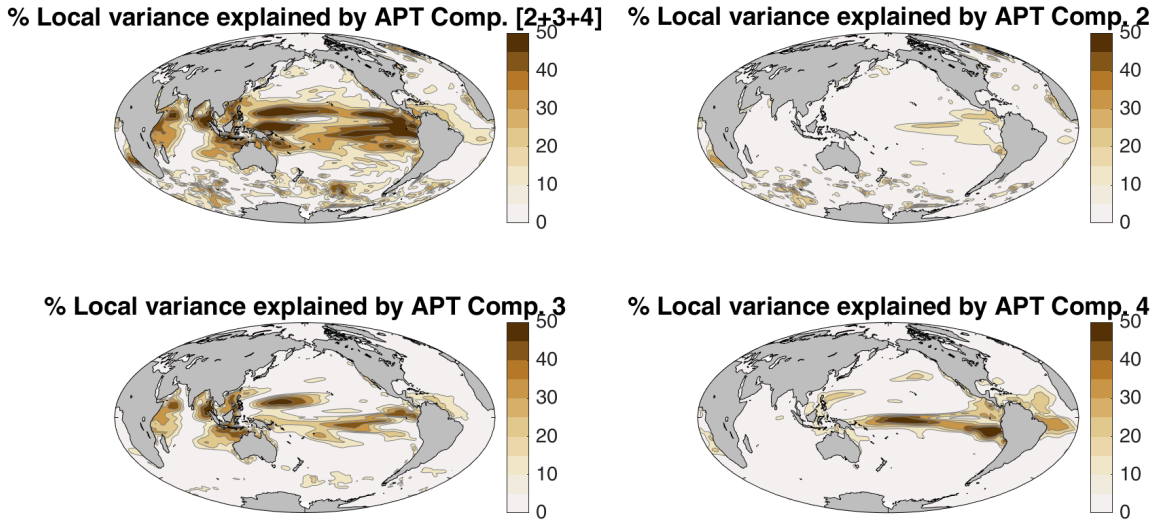


Additional figures showing the three leading predictable components from APT analysis using the detrended anomalies.

The predictive skills are attributed to the ENSO, EBC and SO variations through the high correlation with PrC3 and PrC2 from APT analysis. However, the explained variances of PrCs are in general very low: “PrC2 only explain 3.6% of the total variance”; “PrC3 and PrC4, together explaining 7.2% of the variance”. How to interpret the predictions with a trivial variance explained? Are the attribution and relationship robust?

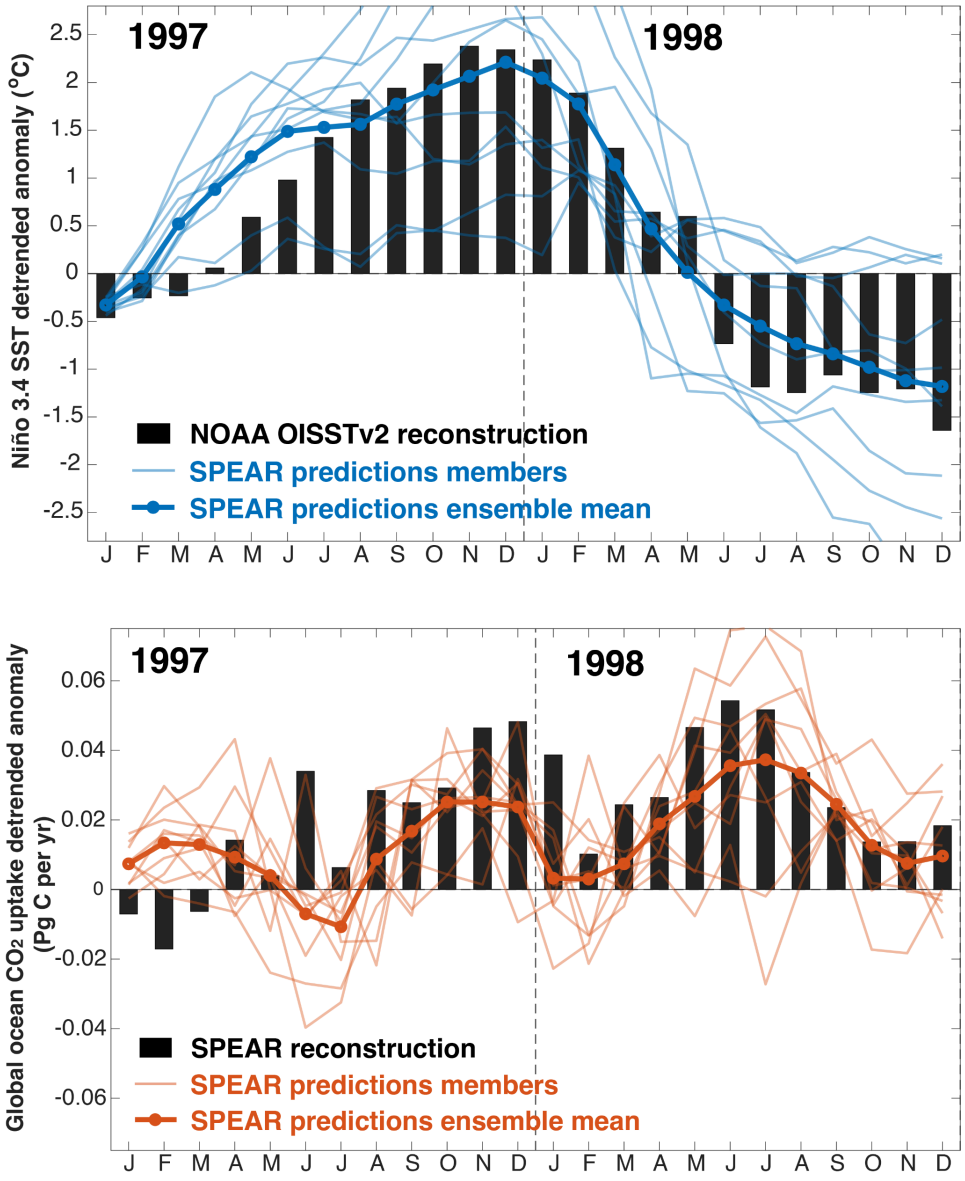
Ensemble forecasts contain substantial ensemble noise, externally forced signals (from CO_2 and aerosol forcing), and predictable natural variability. While predictable components PrC2, PrC3, and PrC4 account for a small fraction of total global variance, their local contributions are substantial. Their local significance can be demonstrated by mapping the regional variance explained by individual APT components (e.g., Yang et al., 2015, Figure 4b).

Figure below shows that the leading predictable components (2-4; excluding component 1 explained by forced variability) explained 30-50% of the local variance at the gridpoint level in the low latitudes, while the local variance in the high latitudes is mostly forced or ensemble noise. This analysis will be included in the Supplemental Materials to hopefully provide further information on the robustness level of the attribution analysis.



ENSO itself has predictive skill on seasonal time frame. How would the relatively short predictive skill of ENSO maintain a predictive skill of air-sea CO2 fluxes of 2 years and longer?

Predictive skill for CO_2 fluxes at a 2-year lead time stems primarily from the delayed but robust response of the ocean carbonate system to non-thermal ENSO impacts. While SST anomalies associated with ENSO typically peak around November–December, non-thermal responses, such as thermocline adjustments and upwelling changes, frequently peak in the following spring and summer. Biological mechanisms further delay the carbon response; for instance, Lim et al. (2022) documented a pronounced post-El Niño chlorophyll rebound in the summer following an event, driven by eastward subsurface iron transport and enhanced dust deposition. Thus, although atmospheric ENSO phenomena have seasonal predictability, predictions initialized in January often successfully capture ENSO evolution during Year 1 and the resulting lagged CO_2 flux response in Year 2, extending skill out to two years. Figure 9a illustrates this lagged relationship, and below we provide additional temporal evaluations of CO_2 flux responses using the 1997–1998 El Niño event as an example. Figure below shows a positive detrended anomaly in air-sea CO_2 flux in 1998 following a positive SST detrended anomaly most prominent in 1997 in the Niño3.4 region. In this case, the CO_2 flux anomaly at a lead time of 2 years (1998) was successfully captured in the predictions initialized in January 1997.



Specific comments:

L62: “of how the”->”of how effective the”
L373: “4i”->”4”
L385-387: Figure 3 capture is inconsistent with the figures shown below
L395: “(e-j)” -> “(f-j)”
Figure 5: The numbers shown are inconsistent with Fig. 2 & 3, in which the predictive skill are shorter than 5 years in most areas. There are more places with 5 year predictive skill in Fig. 5 than those shown in Fig. 2f and Fig. 3f.
L545-546: “Note that this ENSO association explains more variance than PrC2, despite shorter APTs.” This sentence is unclear.
L561-567: the font here is different than the other text.

We thank the reviewer for pointing out these editorial errors and will carefully address them in the revised manuscript.

L573-575: would extended reconstructions and initialized simulations help in this regard on understanding predictability with multidecadal time scale?

We agree with the reviewer on this point. Unfortunately, this study is limited to roughly 40–45 years of historical reconstruction and initialized hindcasts starting in the 1980s. As noted in Lines 647–651, a caveat of relying solely on atmospheric restoring to drive the reconstruction is that our model required an extended spin-up period to synchronize its simulated climate phase with observations following the onset of restoring (1958 in the case of JRA55 forcing). In regions such as the Southern Ocean, this synchronization period lasted 25–30 years. For this reason, we selected 1982 to begin the initialized hindcasts. As we develop refined strategies to properly initialize the ocean interior, and leverage additional reanalysis products with longer records (e.g., ERA5), we expect this synchronization period to be substantially shortened. This will allow us to extend our hindcasts and improve our understanding of multidecadal predictability.
