Implementation and validation of a supermodelling framework into CESM version 2.1.5

William E. Chapman [1], Francine Schevenhoven [2], Judith Berner [1], Noel Keenlyside [2,3], Ingo Bethke [2], Ping-Gin Chiu [2], Alok Gupta [3], and Jesse Nusbaumer [1]

[1]National Center for Atmospheric Research, Boulder, Co. USA
[2]Geophysical Institute and Bjerknes Centre for Climate Research, University of Bergen, Bergen, Norway
[3]Nansen Environmental and Remote Sensing Center, Bergen, Norway

Correspondence: William E. Chapman (wchapman@ucar.edu)

Abstract.

Here we present a research framework for the first atmosphere-connected supermodel using state-of-the-art atmospheric models. The Community Atmosphere Model (CAM) versions 5 and 6 exchange information interactively while running, a process known as supermodeling. The primary goal of this approach is to synchronize the models, allowing them to create a new dynamical system which can theoretically benefit from each component model, in part by increasing the dimensionality of the system.

In this study, we examine a single untrained supermodel where each model version is equally weighted in creating pseudo-observations. We demonstrate that the models synchronize well without decreased variability, particularly in storm track regions, across multiple timescales and for variables where no information has been exchanged. Synchronization is less pronounced in the tropics, and in regions of lesser synchronization we observe a decrease in high-frequency variability. Additionally, the low-frequency modes of variability (North Atlantic Oscillation and Pacific North American Pattern) are not degraded compared to the base models. For some variables, the mean bias is reduced compared to control simulations of each model version as well as the non-interactive ensemble mean.

1 Introduction

Climate models are essential for understanding and analyzing the complex dynamics of our Earth system. However, significant uncertainties remain, primarily due to the challenges in accurately parameterizing key processes and the biases inherent in different components of these models. These biases often exceed the projected climate change signals and the natural background variability that we aim to predict (Palmer and Stevens, 2019). Numerous options to improve climate representation are actively being explored, including enhancing subgrid physics, the incorporation of stochastic terms (e.g., Berner et al., 2008, 2012, 2017), utilizing machine-learned parameterizations and closures developed from observations or high-resolution model runs (e.g., Gregory et al., 2023; Chapman and Berner, 2024; Watt-Meyer et al., 2021; Bretherton et al., 2022), and increasing climate model resolution to directly resolve specific processes instead of parameterizing them (e.g., Judt, 2018; Palmer, 2014; Segura et al., 2025). Often these approaches effectively increase the dimensionality of the prediction system in

some form by adding additional degrees of freedom. However, these methods are challenging to develop and often too computationally intensive to practically implement. Despite these challenges, progress is crucial for improving climate predictions and informing policy decisions on climate adaptation and mitigation. Thus, alternative methods, which rely on the current generation of models, must be tested.

A simple approach to improving model representation is multi-model averaging, performed after individual models have been run. This method has been shown to reduce climate model biases in various applications (e.g., North American Multi-Model Ensemble, Coupled Model Intercomparison Project). These non-interacting ensembles (NIE) reduce errors by balancing the biases from multiple models. Moreover, advanced ensemble weighting schemes can further improve NIE effectiveness (e.g., Weigel et al., 2010; Tegegne et al., 2019). However, NIEs have limitations because they cannot combine model outputs in real-time, making them confined to the attractor space of each individual model. Additionally, biases that are shared across the individual models in an NIE cannot be corrected due to the linear nature of post-process averaging.

Supermodels are designed to address this limitation by creating a new synchronized dynamical system, which consists of the individual models interacting during run time by exchanging either state or tendency information. Since the models exchange information at runtime, the interactive ensemble is effectively of higher dimensionality. Taking advantage of model diversity compensates for individual model bias errors and allows more complex dynamical behavior. However, this is often evidenced in representations of localized structures, rather than in reductions in mean squared error (Duane and Shen, 2023). Supermodeling is a generalization of the interactive ensemble approach introduced by Kirtman and Shukla (2002), who coupled multiple realizations of the same atmospheric general circulation model to a single ocean general circulation model through averaging each models' air-sea fluxes. Since then a number of efforts have focused on linking increasingly complex models from low-dimensional simple models to models of intermediate complexity (van den Berge et al., 2011; Duane et al., 2009, 2018; Schevenhoven et al., 2023), and a framework for a state-of-the-art ocean-connected supermodel has been developed (Counillon et al., 2023).

Supermodels depend on two key principles: firstly, the synchronization of different models rooted in the concept of chaos synchronization in non-linear dynamical systems (Duane and Tribbia, 2001; Pecora et al., 1997); and secondly, the diversity among models can reflect the actual behavior of the target dynamical system. The models can either be directly linked to each other via a subset or all variables (e.g., van den Berge et al., 2011) or connected to their weighted average (e.g., Schevenhoven, 2021; Wiegerinck et al., 2013). The weighted average is also referred to as pseudo-observations, a terminology which will be explained below. A central point is that if the models are sufficiently synchronized the supermodel should not suffer from a decrease in variance. A decrease of variance in a particular region or variable is a sign of an insufficient synchronization and might require the exchange of more information between the interactive ensemble members (Wiegerinck et al., 2013).

To achieve optimal performance, a supermodel must be trained using data from the "truth" such as observations or a reference model. During the training phase, the supermodel interaction coefficients are optimized to formulate the supermodel with the best skill. Since only the interaction coefficients need to be learned, the training effort is substantially less than that used by modern machine learning approaches which learn the entire forward operator of a model (e.g., Watt-Meyer et al., 2023). Efficient methods have been developed to train supermodels (Schevenhoven and Selten, 2017) and have been shown in a

coupled model of intermediate complexity, SPEEDO, connected via the atmospheres only (Severijns and Hazeleger, 2010), these training methods showed promising results (Schevenhoven et al., 2019), even when the observations were sparse and noisy (Schevenhoven, 2021). As the goal of this manuscript is to describe the modeling framework itself, we will henceforth focus on results from an untrained supermodel.

Exchanging information between models at runtime introduces substantial technical complexity and difficulties. Conceptually, each model has to be integrated forward for a number of model timesteps and then paused. Then, information between the component models is exchanged. For simple models all model variables will be available on a single processor and the exchange of information is straight forward. However, for more complex models which need to run on large distributed memory systems, this information sharing can be more difficult. Often, input/output files need to be written, combined, and each component model has to be 'restarted' or 'resumed' after reading in exchanged information. This methodology is reminiscent of the traditional data assimilation (DA) approach, which attempts to synchronize a numerical weather forecast (NWP) with the observed atmospheric state. As such traditional data assimilation in NWP can be considered a special case of supermodeling. This analogy explains why the exchanged information is also called "pseudo-observation".

Duane et al. (2006) recognized this analogy and suggested the use of traditional data assimilation tools (e.g., Du and Smith, 2017) to ingest the pseudo-observations into the components models. This approach was adopted by Counillon et al. (2023) to successfully connect multiple ocean models. However, the overhead of writing and reading input/output files and restarting the component models is computationally very inefficient and prohibits interaction at every timestep. Furthermore a workflow manager is necessary to create the pseudo-observations and delay the restart until the latter are available, especially if one of the component models is slower then the other. Additionally, while some models have pause/resume capability and can be restarted quickly, others, like the one used in this study, take a relatively long time to re-initialize the model, which makes the DA approach for creating a supermodel undesirable.

Here, we describe the technical details of the implementation of the so-far most complex supermodel in a heavily parallelized High Performance Computing environment. It connects two versions of the Community Atmosphere Model (CAM), which is the atmospheric component of the Community Earth System Model CESM (Danabasoglu et al., 2020). The advantage over previous implementations come in through three new developments: 1) the availability of a newly developed Python-Fortran-Bridge in CESM, 2) the adaptation of the existing nudging toolbox (Chapman and Berner, 2023) for our purposes, and 3) the submission of multiple jobs through a single PBS or SLURM scheduler, which allows both component models to get into the same queue.

A newly developed Python-Fortran-Bridge is now available in CESM which allows calls to python routines from the Fortran executable at runtime. Called via the CESM-internal workflow, the python-calls take the role of the workflow manager, control the generation of the pseudo-observations and effectively introduces pause/resume functionality into CESM.

The CESM nudging toolbox was used by Chapman and Berner (2024) to compare nudging tendencies to DA increments. In the supermodeling context, its infrastructure can be easily adapted to facilitate nudging to the (weighted) model-average as well as specifications for the interaction interval and which variables should be connected via Fortran namelist parameters.

While we present only first results using an unweighted, atmosphere-connected supermodel, we stress that the interoperability of the implementation will make extensions to include ocean-connections straight forward. It is also easy to add additional component models, as long as they are available on the same supercomputer. We also use the above-described training methods to optimize the performance of the supermodel, the results of which will be described in a forthcoming manuscript.

The manuscript is structured as follows: Section 2 describes the components models, verification datatsets and implementation of the supermodel, Section 3 discusses model synchronization and presents results supporting a successful implementation. Section 4 provides a discussion and concludes with the findings.

2 Implementation and Synchronization Methodology

2.1 Component Models and experiment setup

We combine different simulations of the Community Atmosphere Model (CAM), an atmospheric general circulation model (AGCM) developed at the National Center for Atmospheric Research with extensive community support. Our supermodel integrates CAM version 5 (CAM5; Neale et al., 2010) and CAM version 6 (CAM6; Bogenschutz et al., 2018; Gettelman et al., 2018), each incorporating different physics suites while using the same finite-volume (FV) dynamical core. CAM5 is released as the atmospheric component of CESM1 and CAM6 of CESM2, respectively.

The CAM5 simulation is run from the CAM6 code base with the CAM5 physics flag activated, which configures CAM to specifically use the physics schemes from CAM5.1 (CESM1.0.6). CAM5.1 treats stratiform cloud microphysics with a two-moment formulation (Morrison and Gettelman, 2008). The spatial distribution of shallow convection is simulated with a set of realistic plume dilution equations (Park and Bretherton, 2009). The ice cloud fraction scheme allows supersaturation via a modified relative humidity over ice and the inclusion of ice condensation amount (Gettelman et al., 2010). Descriptions for all other physics schemes (deep convection, PBL, radiation, etc.) can be found in Neale et al. (2010).

CAM6 uses the publicly released version of cam_cesm2_1_rel_60 from CESM2.1.5. Significant changes from CAM5 physics include substantial modifications to every atmospheric physics parameterization except for radiative transfer. The Cloud Layers Unified by Binormals (CLUBB, Golaz et al., 2002; Bogenschutz et al., 2013) scheme replaces CAM5 schemes for boundary layer turbulence, shallow convection, and cloud macrophysics. Additionally, an improved two-moment prognostic cloud microphysics (MG2 Gettelman and Morrison, 2015) was introduced between versions. The deep convection parameterization (Zhang and McFarlane, 1995) has been significantly retuned to increase sensitivity to convective inhibition. Both subgrid orographic drag calculation schemes have undergone substantial modifications. The orographic gravity wave scheme now incorporates topographic orientation (ridges) and low-level flow blocking effects. Finally, the previous parameterization of boundary layer form drag, known as turbulent mountain stress (TMS), has been replaced by the scheme of Beljaars et al. (2004).

While our supermodeling implementation utilizes interpolation routines to support different vertical and horizontal resolutions, we use here the resolution for which the atmospheric component models were scientifically released, namely a grid-size of 0.9°N x 1.25°E in the horizontal and 32 hybrid sigma-pressure levels up to 2.26 hPa in the vertical.

The model simulations followed the protocol of the Atmospheric Model Intercomparison Project (AMIP) and are forced by observed monthly sea surface temperatures and sea ice from 1979 to 2005 (26 years), with values linearly interpolated at each time step. The simulations also include prescribed evolutions of aerosol emissions and trace gas concentrations (including CO_2).

2.2 Validation Datasets

We verify the model against the $\sim$0.25° ERA5 reanalysis product (Hersbach et al., 2020) for all fields except precipitation, which is verified against the 1° NOAA GPCP product (Adler et al., 2003). For verification, the ERA5 product is bi-linearly interpolated to the native CAM grid prior to any metric calculation. The GPCP product is regridded to the native CAM grid using a conservative mapping method.

2.3 RMSE and Bias Calculation

As in the NCAR Atmospheric Modeling Working Group Diagnostic Package AMWG (2022), model error is calculated as the sum of the cosine- latitude weighted, root-mean-squared error (RMSE) of the spatial field after a seasonal, monthly, or daily mean has been computed. RMSE was used so that opposite-signed local biases do not cancel and erroneously inflate skill. Percent improvement is determined by first calculating global/regional RMSE and then calculating the percent change of RMSE compared to the reference.

2.4 Super Model Implementation

Our first attempt at implementing a supermodeling framework followed previous work (Counillon et al., 2023) and utilized a workflow manager, CYLC, together with tools from the data assimilation research testbed (DART Anderson et al., 2009) to restart the component models after their interaction via nudging to averaged output files. CESM-specific bottlenecks were 1) the time needed to re-initialize CESM after a restart, since the current CESM version does not have pause/resume capability and 2) re-entering the system queue after each interaction interval. Due to these inefficiencies completing a single year's simulation took approximately 24 hours (at 6-hour coupling), a cost that becomes untenable for multi-year simulations and made it impossible to increase the interaction frequency.

To overcome these difficulties, we devised a new custom workflow management system that eliminates the need for halting and re-initializing the model with each interchange of information between models by employing a `PAUSE/RESUME` mechanism. At the beginning of the physics timestep, the first component model outputs the model state variables—Zonal wind (U), Meridional wind (V), Temperature (T), and Specific Humidity (Q)—and initiates a model pause by writing a `PAUSE` file. Subsequently, CAM calls a Python script that waits for the second model to reach the beginning of its physics timestep and then combines the outputs from both models at the same timestamp. If the component model grids differ, Python interpolation routines are invoked to ensure consistency. Once the output has been processed, the Python script removes the `PAUSE` file, allowing the model to resume operation without the need for re-initialization or re-entering the queue. The implementation of

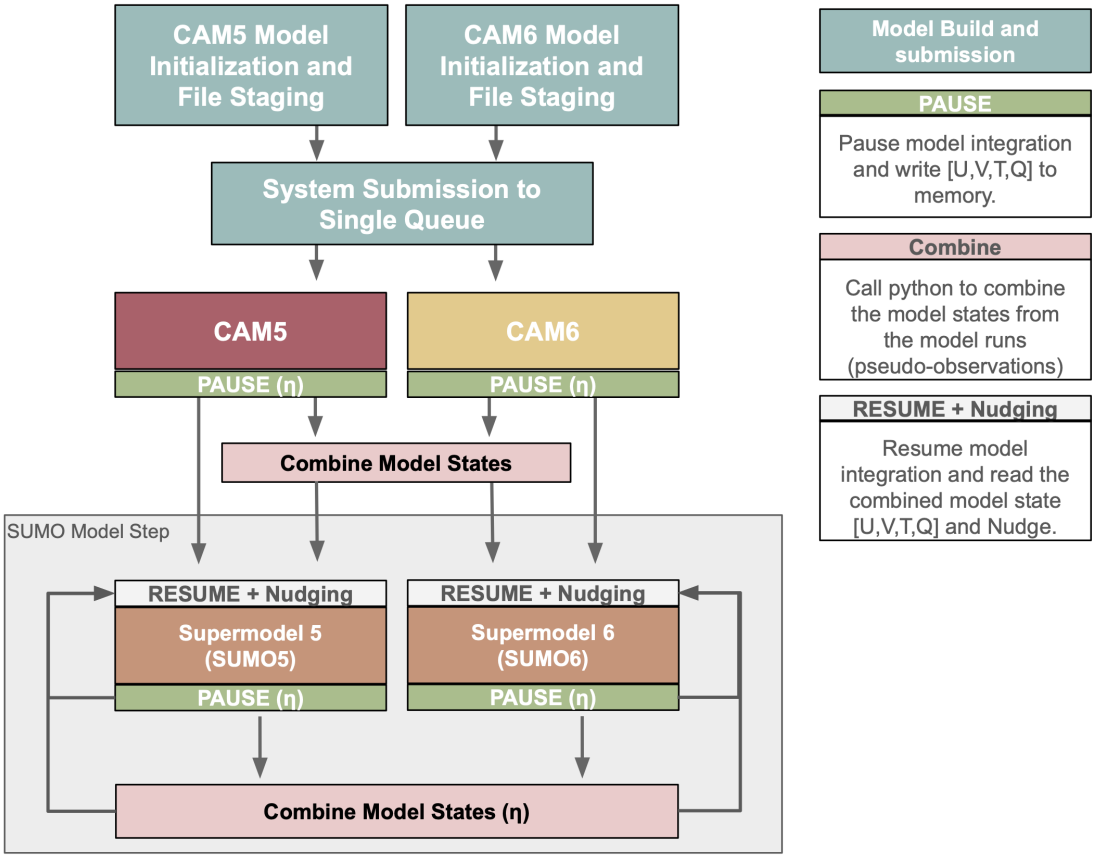

Figure 1. CAM5/6 Supermodel Workflow

efficient mpi-based communication between the models (i.e. standard coupling software) was beyond the scope of the study but is something that should be explored in future efforts.

A nudging tendency is then applied to each component model, following Equation (1), which nudges the model state toward the combined model state (X_{combined}) during the first timestep after the models resume running (e.g., Chapman and Berner, 2024). The user can set the nudging timescale (τ); in this experiment, we use a relaxation timescale of 6 hours. Though we emphasize that the tendency is only applied at the first time-step after the combination (e.g., snapshot nudging).

$$\frac{dX}{dt} = F(X) + \frac{X_{\text{combined}} - X}{\tau} \tag{1}$$

where X is the model state, $F(X)$ represents the model's internal tendencies, and τ is the nudging relaxation timescale.

To address the challenge of submitting multiple jobs through a single PBS/SLURM scheduler, we implemented a batch submission script that allows two model simulations to be launched concurrently while managing resource allocation efficiently.

Our approach ensures that both models utilize the available compute nodes without interfering with each other, thus avoiding scenarios where one model monopolizes the queue while the other remains pending.

Specifically, our submission script:

1. **Prepares model runs** by creating initialization files for both simulations.
2. **Defines model-specific execution settings**, including the number of processing elements required for each job.
3. **Partitions compute resources dynamically** by selecting appropriate node allocations from `$PBS_NODEFILE`, ensuring that both jobs receive the necessary resources without conflicts.
4. **Executes model runs in parallel** using background processes (`&`), allowing both jobs to start simultaneously while still being managed within a single job submission.
5. **Waits for all processes to complete** using `wait`, ensuring that computational resources are fully utilized before job completion.

This method ensures efficient job scheduling and mitigates the risk of asynchronous queuing delays, ultimately reducing computational time.

The resulting CAM5/CAM6 supermodel software workflow diagram is shown in Fig. 1. First, we provide scripts to build, compile, set namelist parameters, and stage necessary python and Fortran files. Then, all component models are submitted to the same submission queue using a single PBS or SLURM scheduler. The component models then run independently using their respective physics package until the first interaction timescale (η) is reached. The interaction timescale η signifies the time after which the supermodels share information. At this point, each model writes output and is paused. Using a python-call from within the CESM workflow of one of the component models, the model states are subsequently averaged to create the combined state or pseudo-observation. The component models then resume, each being nudged towards the pseudo-observations. Once the next interaction timescale is reached, the models are paused and resume after their combined state has been computed. The last two steps are repeated until the desired simulation length is obtained (Fig. 1, grey box).

By adapting the previously developed nudging toolbox (Davis et al., 2022) for our purposes, we can easily set the interaction timescale, interacting variables, pseudo-observation file name, output path locations via namelist parameters.

The CAM5/CAM6 SuperModel, including the CESM Fortran-Python-bridge, supermodel module toolbox with namelist section, and scheduler scripts, are readily available via the GitHub repository. Currently, this system is deployed on two HPC platforms 1) the National Center for Atmospheric Research's Derecho Computer and 2) the Norwegian Research Infrastructure Services' machine, Betzy.

With these improvements, a one-year simulation is now accomplished in 7 hours, and is independent of system queuing time. Moreover, increasing the frequency of coupling does not significantly increase the wallclock time. We acknowledge that this is a significant slowdown from a CAM5/CAM6 simulation which can accomplish a year long simulation in $\sim$ 2.5 hours with identical computational resources.

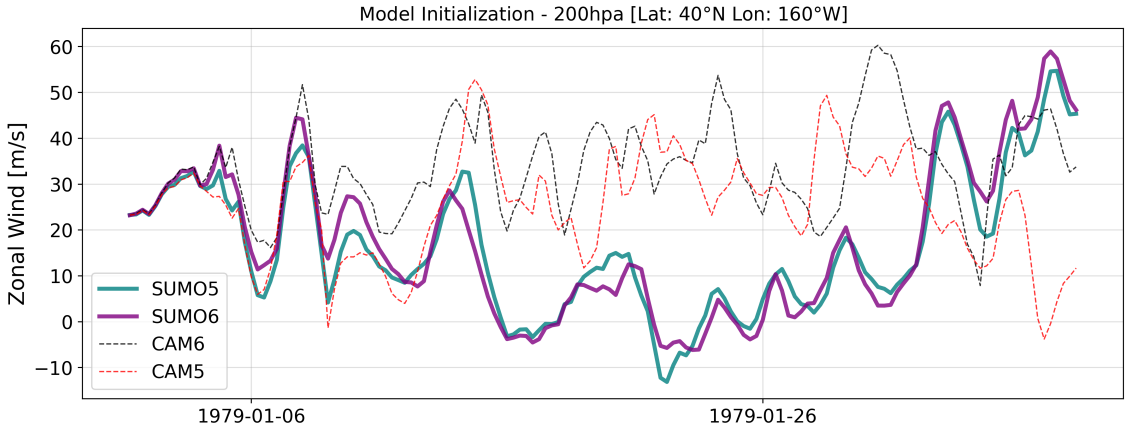

Figure 2. Experiments initialized from the same atmospheric state and integrated for one month. Showing: an independent CAM6 run (dashed black); an independent CAM5 run (dashed red); a supermodel which uses CAM5-physics (SUMO5; teal); a supermodel which uses CAM6-physics (SUMO6; purple) at location [Lat: 40N, Lon: 160W, Lev; 200 hPa].

3 Supermodel Results

We now demonstrate the synchronization and resulting mean state representation for the CAM5/CAM6 supermodel for the period 1979 through 2005. The supermodel uses an interaction timescale of $\eta = 6$ hours and employs snapshot nudging to the unweighted averaged state. In this implementation, the information in U, V, and T are exchanged and nudged while Q is left to evolve freely. We speculate that the main challenges with including specific humidity (Q) in the nudging process stem from the intrinsic properties of moisture in the atmosphere and its coupling with cloud and precipitation processes. This was done because previous work indicated difficulty when adjusting specific humidity Q in CAM in both, nudging (e.g., Chapman and Berner, 2024) and full DA experiments (Raeder et al., 2021).

We show results for four experiments: CAM5, CAM6, the supermodel which uses CAM5-physics, but is nudged to the combined state, SUMO5, and the supermodel which uses CAM6-physics, but is nudged to the combined state, SUMO6. We analyse the 6-hourly averaged prognostic state variables (U, V, T, Q, Surface Pressure (PS)) and standard CAM output which is averaged monthly.

3.1 Synchronization

Figure 2 shows the zonal wind (U wind) at 200 hPa for four experiments started from the same initialization at a single model point ([40°N, 160°W]). The results indicate that while CAM5 and CAM6 vary independently, the two supermodel trajectories synchronize after ca. 15 day and the trajectories stay closely linked throughout the model run.

Figure 3 illustrates the anomaly Pearson correlation coefficients for four atmospheric variables: (U,V,T,Q) at a 6-hourly averaged temporal frequency. These correlations are computed between the two super models (SUMO5, SUMO6) and between

the individual CAM6 model and SUMO6 in model year 1979-1980. The analysis covers the 200hPa level at every grid point. To avoid anomalously high correlations for areas where the variability is largely driven by the annual cycle, we remove the 30-day centered rolling mean of the data. Figure 3 provides a quantitative measure of the degree of similarity between the two supermodel versions and the similarity between the one of the component models, CAM6, and the associated supermodel using also CAM6-physics. It highlights the effectiveness of the supermodeling approach in synchronizing the atmospheric state across different variables.

As to be expected there is no synchronization of the CAM6 and SUMO6 model states, as expressed by a low correlation coefficients of <0.15 for all almost all gridpoints (Fig. 3, right column). Correlations between the SUMO5 and SUMO6 experiments are much higher (Fig. 3, left column) indicating that the synchronization is evident not only across the component models, but also between the supermodels using different physics packages. Synchronization is strong in the U, V, and T fields poleward of 15°, especially in the storm track regions. The Maritime continent region (Lat: [15°S,15°N], Lon:[60°E,200°W]) displays the least amount of synchronization. Q displays less synchronization (Fig. 3g) with lowest correlation in the tropical belt, but still more than the correlation between CAM5 and CAM6. The supplemental material shows the same analysis for pressure levels of 750 hPa and 900 hPa (Fig. 1S and Fig. 2S). Generally, we see greater synchronization at of U,V, & T at higher pressure levels, while Q has a greater synchronization nearer to the surface, which could be a result of a similar sea surface temperature field between the two models.

If the component models are not sufficiently synchronized, the combined model state will exhibit diminished high-frequency variance compared to the individual models. This variance deflation occurs because the supermodel, representing a weighted average, tends to smooth out discrepancies between the models. As a result, the supermodel may lose critical variance, leading to reduced accuracy in capturing fine-scale variability. This issue is structurally related to the double penalty problem in modern machine learning for Numerical Weather Prediction (Brenowitz et al., 2024).

Poor synchronization between models, whether spatial or temporal, leads to an averaging effect that disproportionately smooths out high-frequency variations, dampening the system's true variability. Studies show that the less synchronized the models, the more the supermodel's variance is compromised by this effect, detracting from its ability to capture dynamic processes accurately (Counillon et al., 2023).

To examine the supermodel for signs of significant variance deflation, we compare histograms of 6-hourly averaged wind speed values (Fig. 4). We note that the CAM5 (red-dashed) and CAM6 (black-dashed) distributions are quite similar, this is likely do to the model tuning activity at NCAR prior to the model release. We detect a slight damping of the background winds near the mode of the distribution, but no degradation of the highest wind speeds. Overall, the difference between the component and super-models is minimal.

Even if the full fields do not suggest variance deflation, previous work using nudging (e.g., Chapman and Berner, 2024) suggests that any linear relaxation back to some sort of reference field is expected to reduce band-pass filtered variance. Hence we compute the standard deviation of the 12h to 5d band passed-filtered winds for U and V at 200 hPa (Figure 5). The zonal average (line), and zonal standard deviation (shading) are shown for SUMO6 (teal) and CAM6 (black) in the right hand column for the U and V winds (Fig. 5c & f, respectively). There is a significant damping of variability in this frequency band in areas

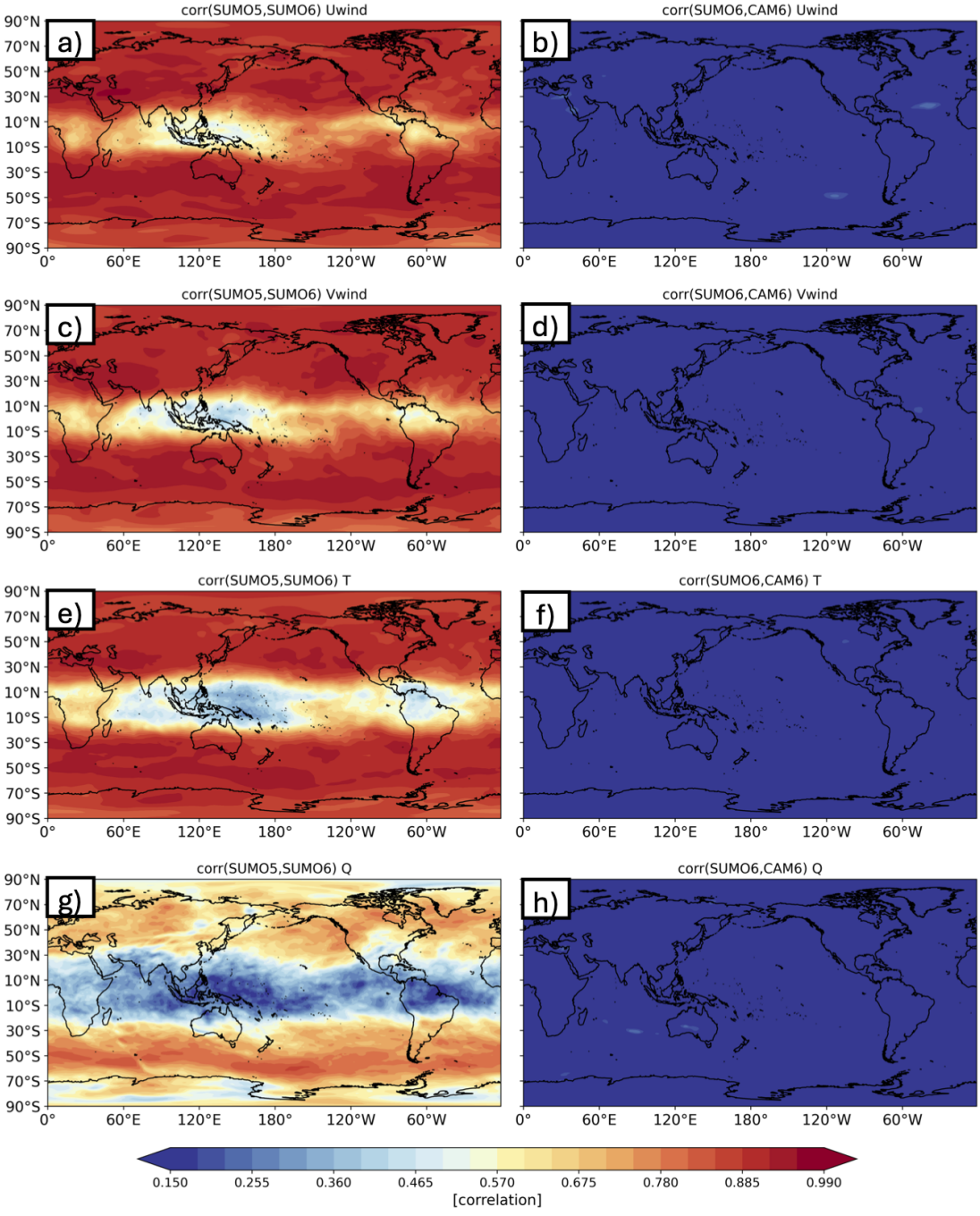

Figure 3. Anomaly correlation between the SUMO5 and SUMO6 s (left: a,c,e,g) and between the SUMO6 and CAM6 experiment (right: b,d,f,h). Shown are 6hourly-averaged model variables zonal wind (a,b), zonal wind (c,d), temperature (e,f), and specific humidity (g,h) at 200 hPa for the period spanning 1979-1980. Anomalies are computed by removing a 30-day centered mean at every timestep.

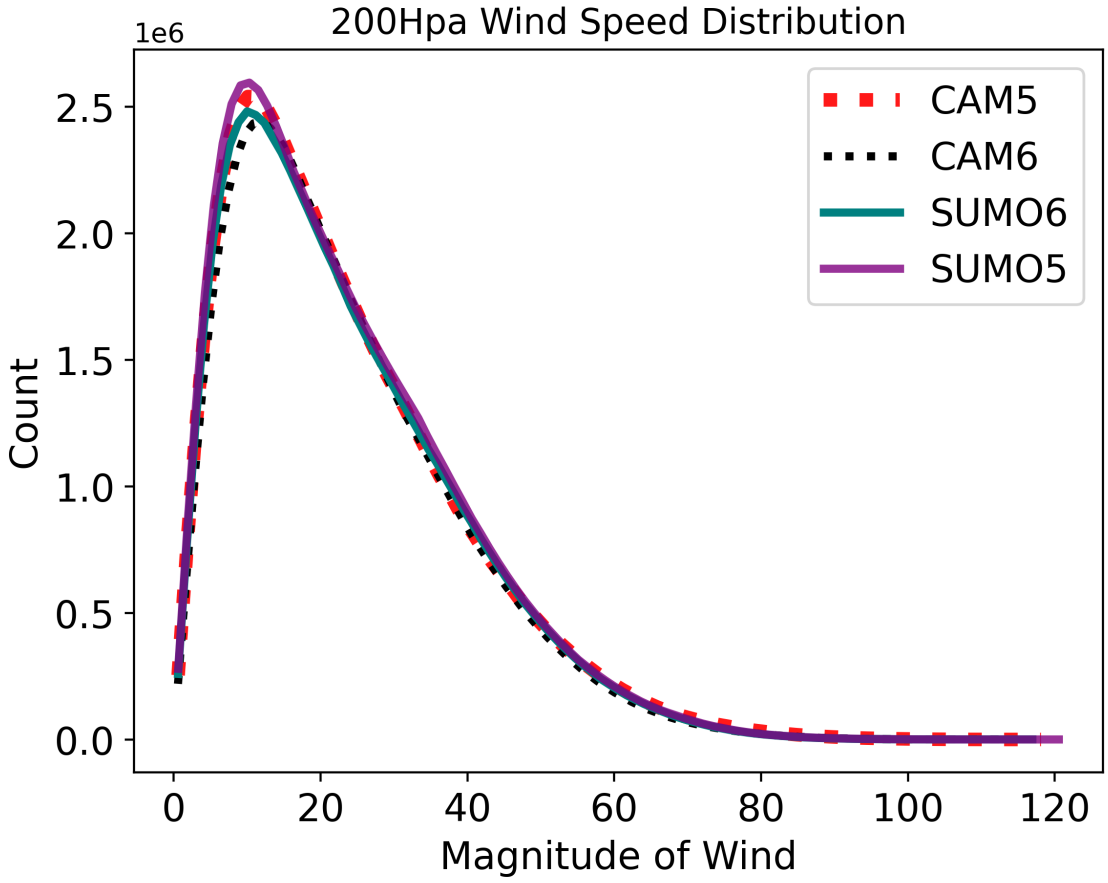

Figure 4. Histogram of all 6-hourly averaged wind speed values in model years 1979 at 200 hPa. Showing: an independent CAM6 run (black); an independent CAM5 run (red); the CAM5 supermodel (teal); CAM6 supermodel (purple).

where we find lower model synchronization like the tropics and the poles (Fig. 3a & c), especially over the maritime continent. In nudging studies, moving to an observations frequency of less than 6 hours seems to alleviate effects of damping (Davis et al., 2022), so we hypothesize that increasing the SUMO interaction frequency will be beneficial with regard to minimizing the damping of high frequency variability.

3.2 Low Frequency Modes of Variability

Evaluating the performance of an atmospheric model requires the adequate depiction of natural climate variability and significant low-frequency climate modes (e.g., Phillips et al., 2014). Intraseasonal variability arises from complex dynamical processes operating across multiple timescales, which subsequently influence downstream weather patterns (e.g., Branstator, 1992; Simmons et al., 1983; Wallace and Gutzler, 1981). The model's background climatology significantly influences this low-

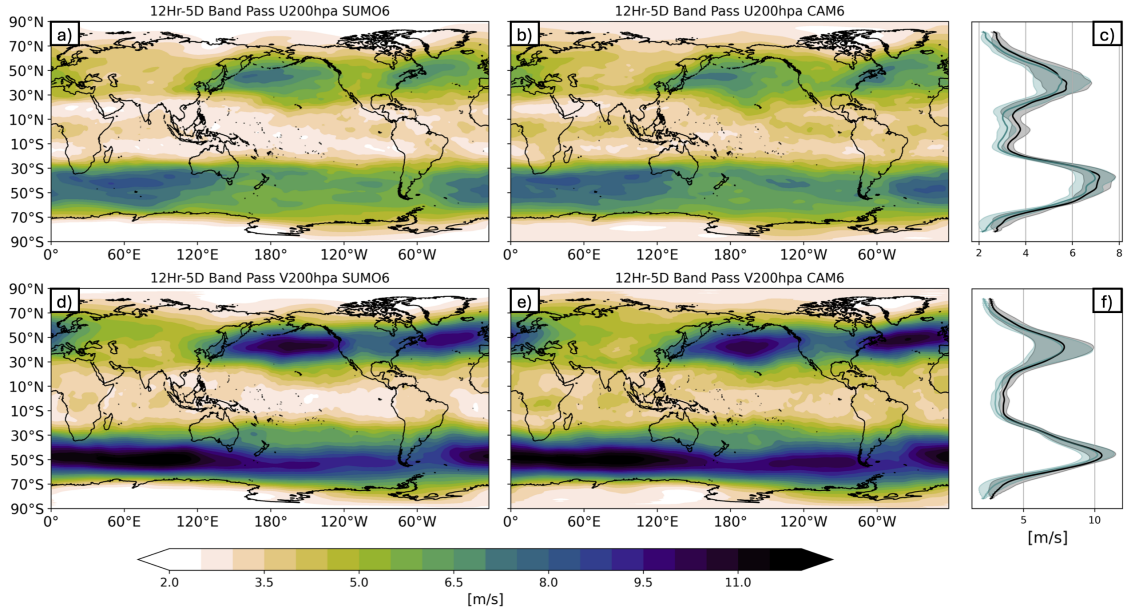

Figure 5. Standard deviation of 12 hour to 5 day band passed U (a,b) and V (d,e) winds for the SUMO6 (column I) and CAM6 (column II) model runs. And, the zonal average (line) and standard deviation (shading) of the the band passed winds for U (c) and V (f) for the SUMO6 (teal) and CAM6 (black) runs.

frequency variability, with several mechanisms proposed for its sustenance and growth. These mechanisms include the development of low-frequency anomalies due to instabilities in the zonally asymmetric midlatitude jet (e.g., Branstator, 1990, 1992; Frederiksen, 1983; Simmons et al., 1983), alterations in quasi-stationary eddies linked to changes in the zonal-mean flow (e.g., Branstator, 1984; Kang, 1990), tropical heating or orographic forcing (e.g., Hoskins and Karoly, 1981; Sardeshmukh and Hoskins, 1988), and vorticity fluxes from high-frequency eddies (e.g., Branstator, 1992; Egger and Schilling, 1983; Lau, 1988; Ting and Lau, 1993). Accurately representing this low-frequency variability is vital for climate models, as the numerous interactions that contribute to an accurate depiction of low-frequency modes are indicative of the model's reliability.

To extract the leading patterns of variability we perform a empirical orthogonal function (EOF) decomposition on the monthly anomaly fields. The climatology is defined as the monthly mean from the full 26-year run or reanalysis product. All EOF patterns are area weighted by the square root of the cosine (latitude) prior to decomposition. We express the orthogonal spatial field as the pointwise regression of each time series with a one-standard deviation change of the temporal principal component. The DJF Pacific - North American Pattern (PNA, Fig. 6), and North Atlantic Oscillation (NAO, Fig. 3S) are examined in detail. These patterns are defined as in Phillips et al. (2014) (NCAR's Climate Variability Diagnostic Package) as the leading mode of atmospheric variability in the region [20-85°N, 120°E-120°W] and [20-80°N, 90°W-40°E], respectively.

Twenty-six years and a single atmospheric realization is likely too short to adequately assess the spatial bias of either the PNA or the NAO (Deser et al., 2017). Therefore, we simply examine the patterns to show general loading locations (Fig. 6a-d) and

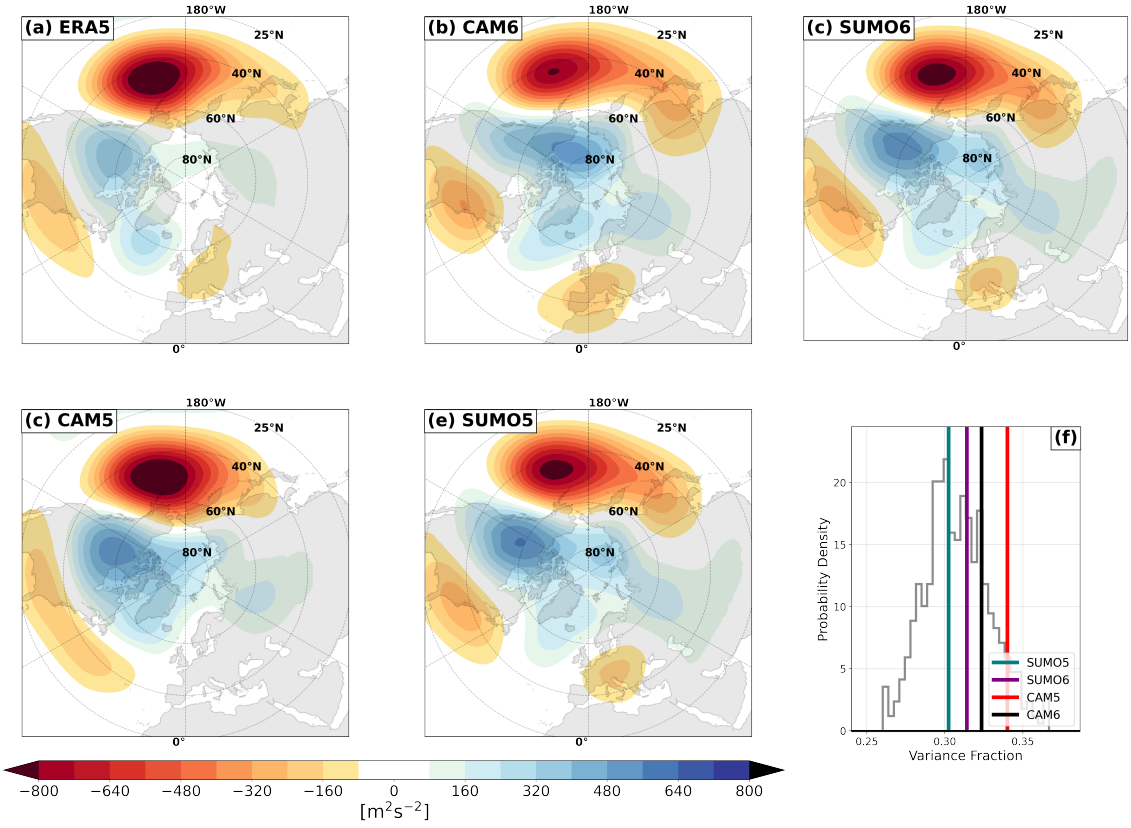

Figure 6. DJF 500 mb geopotential leading mode of variability over the region [NAO, 20-80°N, 90°W-40°E] ERA5 (a), CAM6 (b), CAM5 (c), and SUMO6 (d). Also, the explained variance in each model experiment (solid lines), and the bootstrapped spread of explained variance in the observations (e, grey histogram).

compare the representation of total variance to observations (Fig. 6e, grey histogram). To develop the histogram of explained variances we sub-sample the ERA5 observations into random twenty-six year chunks and bootstrap the EOF calculation 500 times (as in, Chapman and Berner, 2024). The PNA shows a classic stationary Rossby wave pattern spanning from the central Pacific, across Canada and through Florida for every simulation. It is encouraging that the pattern in the supermodels is nearly identical, showing that low frequency modes of variability are also synchronized and that connecting U, V, and T lead to the synchronization of the geopotential height field.

The principal components corresponding to the SUMO5 and SUMO6 PNA exhibit a Pearson correlation coefficient of 0.992, whereas the correlation between the CAM6 PNA PC and the SUMO6 PC is only 0.27. We would expect some correlation due to tropical SST forcing in the AMIP runs (Wallace and Gutzler, 1981). Additionally, the PNA's variance explained sits well within the spread of the observations (Fig. 6e).

The NAO is slightly less synchronized (Fig. 3S) with a Pearson correlation coefficient of 0.75 between the two supermodel runs, but this correlation is still much higher than the correlation of -0.012 between the SUMO6 and CAM6 run.

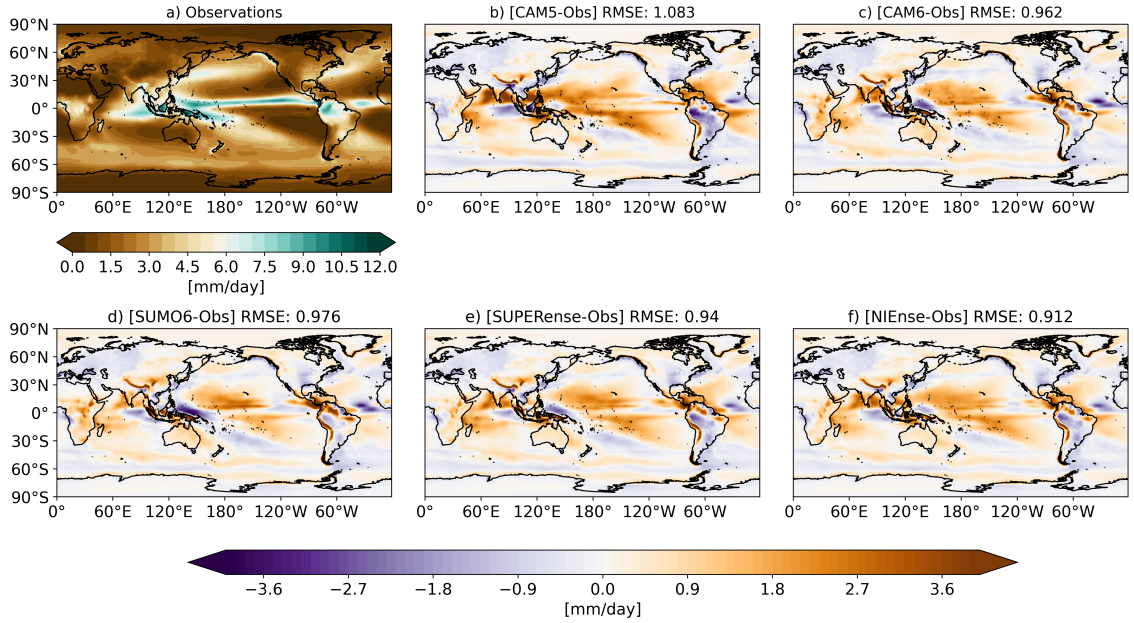

Figure 7. The observed annual precipitation climatology [mm/day] in the NOAA GPCP product (a) and the biases of the CAM5 model (b), CAM6 model (c), SUMO6 (d), SUMO5 (e), and NIEnse (f) relative to the observations. The model RMSE [mm/day] is shown in the title of each panel. The color bars indicate the precipitation bias in mm/day, with orange representing positive bias (too much precipitation) and purple representing negative biases (too little precipitation).

3.3 Impact on mean-field biases

One motivation for developing supermodels is their potential to reduce mean-field biases. In our work, we emphasize the supermodeling implementation connecting CESM components without performing any training—any bias improvements are muted. Therefore we do not anticipate substantial reduction in bias, but there may be minor improvements beyond that from averaging of non-interactive simulations because error compensation at an early stage (Schevenhoven et al., 2023; Duane and Shen, 2023). To ensure that the synchronization did not introduce significant errors or artifacts, we diagnose the climatological biases. For most variables, the SUMO biases fall between those of CAM5 and CAM6 (see supplemental Table 1S for a statistics on the prognostic variables at multiple model levels), a pattern that holds true even when the fields are stratified by season (data not shown).

Figure 7 shows the annual precipitation climatology in the NOAA GPCP product and the model biases (Model - Observations). The SUMO5 and CAM5 precipitation biases are similar, likely because they share the same convection and boundary layer schemes (see section 2.1.1 & 2.1.2). The same is true for the SUMO6 and CAM6 experiments.

In SUMO6 the largest differences from their respective constituent models are over the tropics with loading differences from the Bay of Bengal through the international dateline, and again off of the Pacific Coast of Central America (see Fig. 8a). This

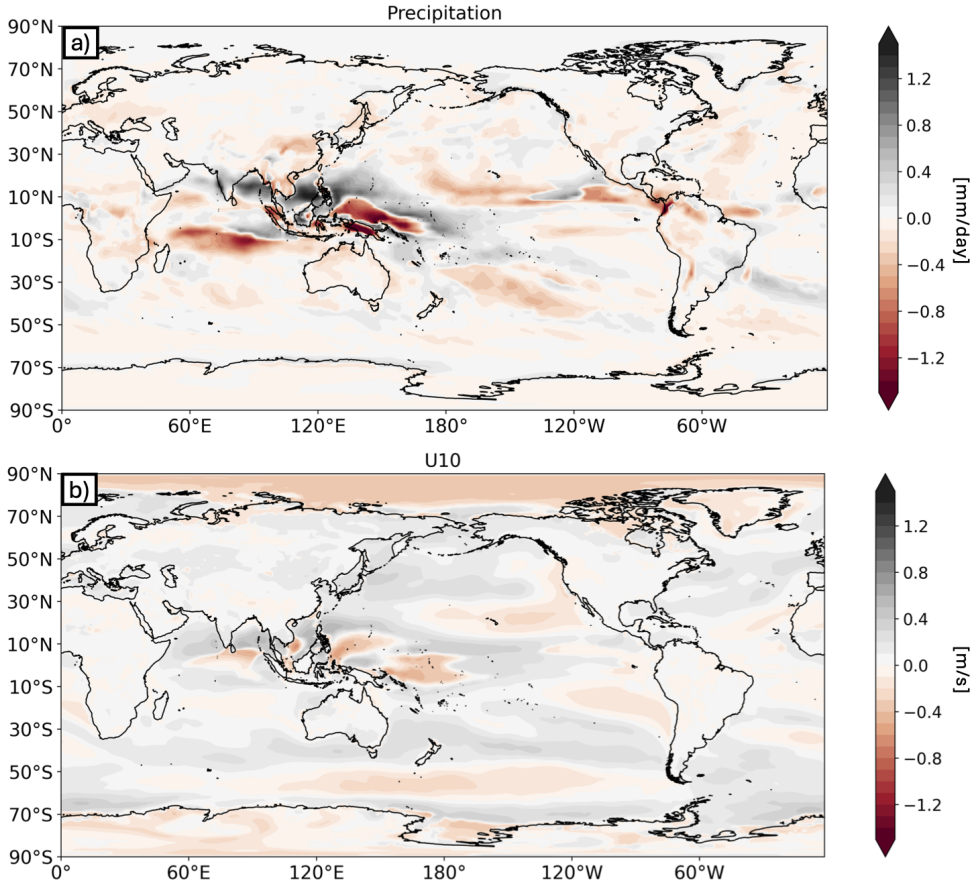

Figure 8. Annual climatological difference between SUMO6 and CAM6 for annual precipitation (a) and annual ten-meter winds (b).

indicates that the synchronization of the prognostic variables is likely affecting the monsoonal regions and deep convective zones.

We introduce a new experiment, the non-interactive ensemble (NIEnse), which is a multi-model ensemble mean of the CAM5 and CAM6 simulation runs, which interestingly has the lowest RMSE with a value of 0.91 mm/day. Figure 9 shows the same analysis but for the 10-meter wind speed. For wind speed, SUMO6 has an RMSE of 0.91 m/s, even outperforming the NIEnse simulation. The largest changes between the respective CAM and SUMO models occur over the maritime continent. This is particularly noticeable in the SUMO5 simulation (see Fig. 9e), which is closer to CAM6 than to CAM5 in this region.

To examine this more closely, we compare the differences between the NIEnse and an ensemble formed as the average of the two SUMO runs (SUPERense). Figure 10 shows the absolute difference in bias [(|NIEnse-Observations|) - (|SUPERense-Observations|)] for annually averaged precipitation (top) and U10 (bottom). Positive values (green) indicate that the SUPERense is outperforming the NIEnse while negative values (blue) indicate the opposite. It is clear that the SUMOs formed

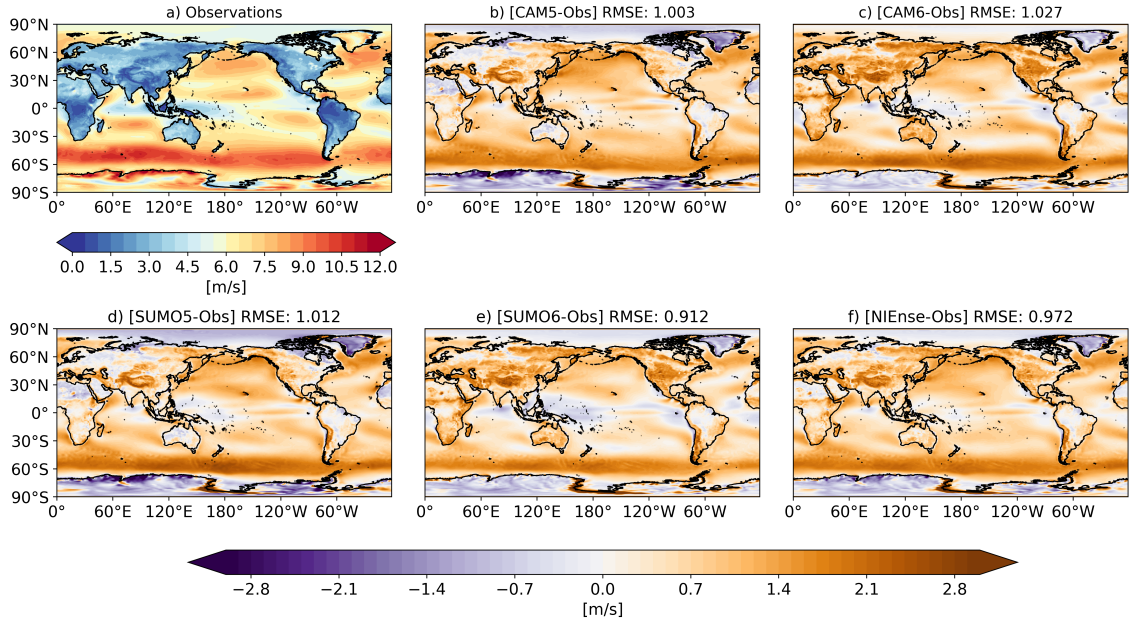

Figure 9. As in Figure 7, but for 10-meter wind speed.

their own dynamical systems with distinct biases. We observe that the SUPERense represents a 5% improvement to RMSE over the NIEnse for annual U10 winds and a 3% degradation of annual precipitation when compared to the NIEnse.

Focusing on the U10 winds (Fig. 10b), the Pacific low-cloud deck regions are attenuated and degraded while the eastern boundary current regions are enhanced (see, for example, the Kurishio and Gulf stream extensions). Unlike, Counillon et al. (2023), we do not find that the areas of low synchronization necessarily lead to areas of high bias.

4 Conclusions

In this work, we give technical details on the implementation of the CAM5/CAM6 supermodel, which is the first to connect two atmospheric components of general circulation models in an HPC setting.

Our implementation leverages three new developments: 1) The exchange of information is managed through a novel Python-FORTRAN I/O interface that avoids the need to stop and start each model. This circumvents in CESM the costly initialization stage and introduces a PAUSE/RESUME capability (Fig. 1). This Python-FORTRAN bridge was also used to manage the timestamps in the output files, and efficiently write the pseudo-observations files. 2) All component models are submitted through a single PBS or SLURM scheduler, which allows both component models to get into the same queue. This minimizes the time one component model has to wait for the other one to finish. Without these two improvements the supermodel would have been too slow to produce multi-year simulations. 3) We were able to adapt the CESM nudging toolbox (Davis et al., 2022) for our purposes, so that we have full control over e.g. which pseudo-observation variables we want to connect.

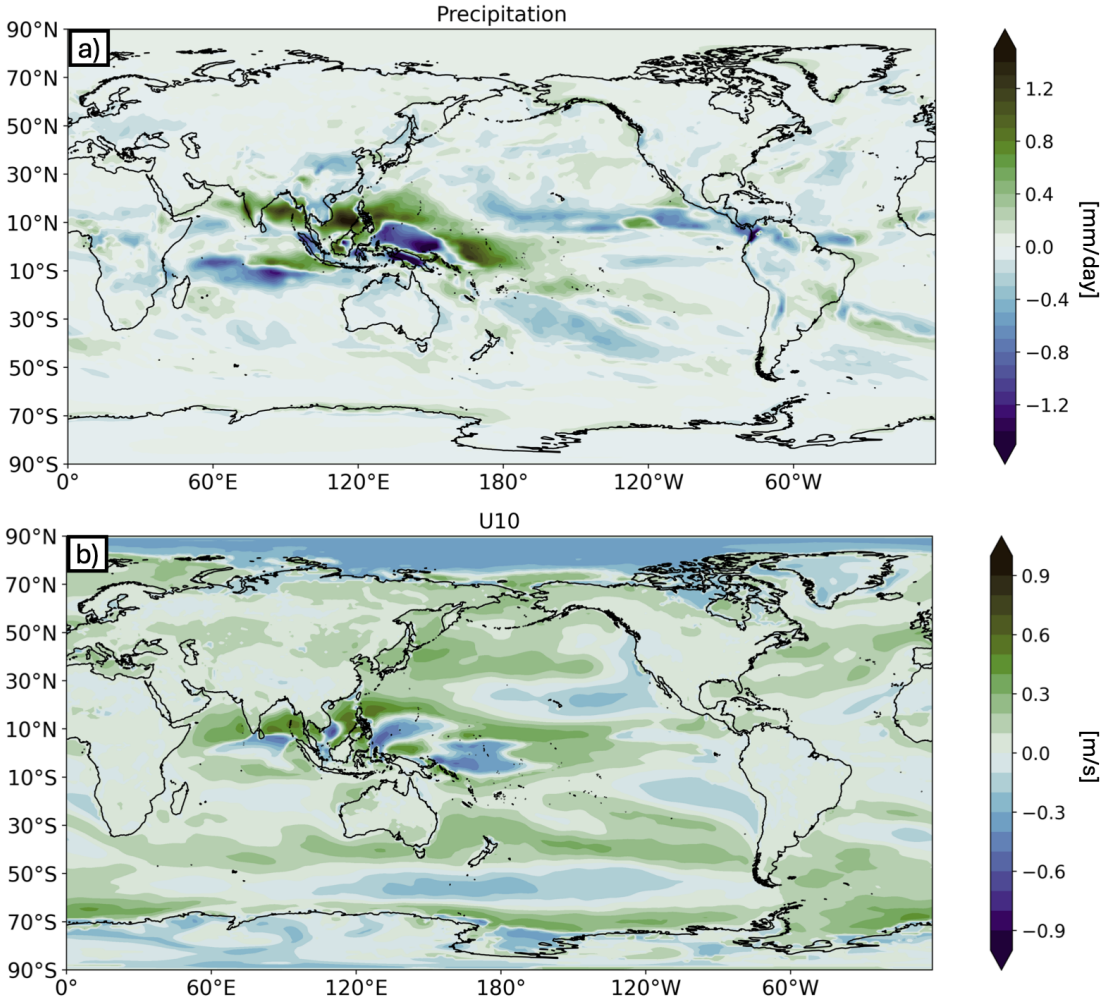

Figure 10. The annual absolute bias difference of NIense and the SUPERense ((|NIEnse-Observations|) - (|SUPERense-Observations|)) for precipitation (a) and 10-meter windspeed (b). Positive values indicate SUPERense is more skillful and vise-versa.

The supermodel framework is readily available for earth system research via our public GitHub repositories, making it relatively easy to port if active CESM systems are installed on a machine. Currently, the framework is available on both the NCAR and Norwegian supercomputers. Our driver scripts set-up the users constituent models and pseudo observations. All software is made available through Github repositories (see the code and data availability section).

To test our implementation we linked the CAM5/CAM6 atmosphere and confirmed that synchronization occurs across various temporal scales and variables. Additionally, even though the supermodels only exchange limited information (U, V, and T) every 6 hours, fields outside of the exchanged information exhibit synchronization across multiple time scales.

A key consideration for future work is assessing how the supermodel maintains physical consistency in terms of energy conservation. While our current analysis has focused primarily on wind and temperature fields, a thorough evaluation of radiative fluxes, surface turbulent fluxes, and the overall energy budget will be essential before extending this approach to coupled Earth system models. Ensuring an accurate energy balance will be crucial for improving the fidelity of the supermodel and avoiding unintended biases in simulations. A potential promising avenue of research would be to dynamically connect the fluxes in our supermodeling state as in Shen et al. (2017) which could ensure that each model's energy fluxes are accounted for in the supermodeling framework.

Additionally, our study has primarily examined large-scale variability modes such as the PNA and NAO. However, given the noted reduction in high-frequency variability over the tropics, it will be important to assess the impact on phenomena such as the Madden-Julian Oscillation (MJO) and convectively coupled equatorial waves. These modes play a key role in tropical variability and global teleconnections, and their representation within the supermodel framework remains an important avenue for future research.

In this study, we only nudge to pseudo-observations which are the equally-weighted mean of the two component models. Unweighted-mean supermodels can lead to partial synchronization regimes and localized variability damping (see also Counillon et al. (2023)). We find that in regions of lesser synchronization, some model variability is damped due to the smoothing effect of averaging over dissimilar fields, though the effects are minimal and should improve as the information exchange frequency (η) is increased.

Since our supermodel is untrained, we do not expect large improvements in the mean-fields biases. However, even with the untrained field, we find evidence of some improvement to the model climatological biases (see Fig. 9). We specifically examined localized structures for signs of improvement (Duane and Shen, 2023), but found no evidence to support any enhancement. With the computational efficiency of our implementation, we can now focus on the question if a trained supermodel of comprehensive Earth-System models can outperform its component models as demonstrated for simpler systems. Since our implementation is using the latest CESM infrastructure, an extension to coupled framework is straight forward.

Additional work will explore the use of machine learning techniques to dynamically optimize the weights and improve the performance of the CESM supermodel.

Code and data availability. To promote transparency and reproducibility, this study includes two code repositories:

1. All figure scripts are readily accessible and can be downloaded using the provided code on Zenodo (Chapman, 2025) to produce all figures.

2. To create all model runs and build your own supermodel, refer to Chapman et al. (2025). This second repository contains the setup for the SuperModel and its constituent models, including source modifications, model build scripts, and namelists for running the described CAM versions.

Comprehensive instructions for each step of this study are documented in the repository's README file. Raw ERA5 Reanalysis data can be obtained from the NSF NCAR Research Data Archive (European Centre for Medium-Range Weather Forecasts, 2019). The Global Precipitation Climatology Project (GPCP) Monthly Analysis Product data is provided by the NOAA PSL, Boulder, Colorado, USA, and can be accessed at https://psl.noaa.gov. CAM5 and CAM6, with directions to run, can be accessed at https://github.com/ESCOMP/CESM

Author contributions. WEC and FS developed the code to build, submit, and integrate the supermodels, led the output model diagnostics, and spearheaded the writing of the manuscript. JB and NK assisted in the interpretation of results, contributed to the writing, and secured project funding. IB, AKG, PC, and JN provided support in software engineering, contributed to the interpretation of results, and assisted in the writing of the manuscript.

Competing interests. The authors declare no competing interests

Acknowledgements. We would like to extend our gratitude to Jim Edwards for his invaluable guidance in software engineering for this project. We thank two anonymous reveiwers and the editor for their input which strengthened the manuscript. This work was supported by NSF project 2015618 - Coherent Precipitation Extremes in a Supermodel of Future Climate, ERC PoC grant number 101101037 and the Impetus4Change EU Horizon Europe project (grant no. 101081555). This research received support through Schmidt Sciences, LLC. We acknowledge discussions on synchronization and supermodeling with Dr. Greg S. Duane and Prof. Jeffrey B. Weiss. We gratefully acknowledge Norwegian HPC resources provided by sigma2 (NN9385K, NS9207k).

References

Adler, R. F., Huffman, G. J., Chang, A., Ferraro, R., Xie, P.-P., Janowiak, J., Rudolf, B., Schneider, U., Curtis, S., Bolvin, D., et al.: The version-2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979–present), Journal of hydrometeorology, 4, 1147–1167, 2003.

AMWG: AMWG Diagnostics Package, NCAR CESM Atmosphere Model Working Group, 2022.

Anderson, J., Hoar, T., Raeder, K., Liu, H., Collins, N., Torn, R., and Avellano, A.: The data assimilation research testbed: A community facility, Bulletin of the American Meteorological Society, 90, 1283–1296, 2009.

Beljaars, A. C., Brown, A. R., and Wood, N.: A new parametrization of turbulent orographic form drag, Quarterly Journal of the Royal Meteorological Society, 130, 1327–1347, 2004.

Berner, J., Doblas-Reyes, F. J., Palmer, T. N., Shutts, G., and Weisheimer, A.: Impact of a quasi-stochastic cellular automaton backscatter scheme on the systematic error and seasonal prediction skill of a global climate model, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 366, 2559–2577, 2008.

Berner, J., Jung, T., and Palmer, T. N.: Systematic model error: The impact of increased horizontal resolution versus improved stochastic and deterministic parameterizations, Journal of Climate, 25, 4946–4962, 2012.

Berner, J., Achatz, U., Batte, L., Bengtsson, L., de la Cámara, A., Christensen, H. M., Colangeli, M., Coleman, D. R. B., Crommelin, D., Dolaptchiev, S. I., et al.: Stochastic parameterization: Toward a new view of weather and climate models, Bulletin of the American Meteorological Society, 98, 565–588, 2017.

Bogenschutz, P. A., Gettelman, A., Morrison, H., Larson, V. E., Craig, C., and Schanen, D. P.: Higher-order turbulence closure and its impact on climate simulations in the Community Atmosphere Model, Journal of Climate, 26, 9655–9676, 2013.

Bogenschutz, P. A., Gettelman, A., Hannay, C., Larson, V. E., Neale, R. B., Craig, C., and Chen, C.-C.: The path to CAM6: Coupled simulations with CAM5. 4 and CAM5. 5, Geoscientific Model Development, 11, 235–255, 2018.

Branstator, G.: The relationship between zonal mean flow and quasi-stationary waves in the midtroposphere, Journal of Atmospheric Sciences, 41, 2163–2178, 1984.

Branstator, G.: Low-frequency patterns induced by stationary waves, Journal of Atmospheric Sciences, 47, 629–649, 1990.

Branstator, G.: The Maintenance of Low-Frequency Atmospheric Anomalies, Journal of the Atmospheric Sciences, 49, 1924–1946, https://doi.org/10.1175/1520-0469(1992)049<1924:TMOLFA>2.0.CO;2, 1992.

Brenowitz, N. D., Cohen, Y., Pathak, J., Mahesh, A., Bonev, B., Kurth, T., Durran, D. R., Harrington, P., and Pritchard, M. S.: A practical probabilistic benchmark for ai weather models, arXiv preprint arXiv:2401.15305, 2024.

Bretherton, C. S., Henn, B., Kwa, A., Brenowitz, N. D., Watt-Meyer, O., McGibbon, J., Perkins, W. A., Clark, S. K., and Harris, L.: Correcting coarse-grid weather and climate models by machine learning from global storm-resolving simulations, Journal of Advances in Modeling Earth Systems, 14, e2021MS002 794, 2022.

Chapman, W.: WillyChap/Chapman_2025_GMD: Figure Release V1 (v1.1.0), https://doi.org/10.5281/zenodo.14983576, 2025.

Chapman, W. and Berner, J.: Deterministic and Stochastic Tendency Adjustments Derived from Data Assimilation and Nudging, Quarterly Journal of the Royal Meteorological Society, 2023.

Chapman, W., Schevenhoven, F., Berner, J., Keenlyside, N., Nusbaumer, J., Bethke, I., Kumar Gupta, A., and Chiu, P.-G.: WillyChap/SuperModel_CAM: PauseResume_v1.1.0, https://doi.org/10.5281/zenodo.14983620, 2025.

Chapman, W. E. and Berner, J.: Deterministic and stochastic tendency adjustments derived from data assimilation and nudging, Quarterly Journal of the Royal Meteorological Society, 150, 1420–1446, 2024.

Counillon, F., Keenlyside, N.-S., Wang, S., Devilliers, M., Gupta, A., Koseki, S., and Shen, M.-L.: Framework for an ocean-connected supermodel of the Earth System, JAMES, 15, https://doi.org/https://doi.org/10.1029/2022MS003310, 2023.

Danabasoglu, G., Lamarque, J.-F., Bacmeister, J., Bailey, D., DuVivier, A., Edwards, J., Emmons, L., Fasullo, J., Garcia, R., Gettelman, A., et al.: The community earth system model version 2 (CESM2), Journal of Advances in Modeling Earth Systems, 12, e2019MS001 916, 2020.

Davis, N. A., Callaghan, P., Simpson, I. R., and Tilmes, S.: Specified dynamics scheme impacts on wave-mean flow dynamics, convection, and tracer transport in CESM2 (WACCM6), Atmospheric Chemistry and Physics, 22, 197–214, 2022.

Du, H. and Smith, L. A.: Multi-model cross-pollination in time, Physica D: Nonlinear Phenomena, 353-354, 31 – 38, https://doi.org/https://doi.org/10.1016/j.physd.2017.06.001, 2017.

Duane, G., Tribbia, J., and Kirtman, B.: Consensus on long-range prediction by adaptive synchronization of models, Geophysical Research Abstracts, 11, 13 324, https://meetingorganizer.copernicus.org/EGU2009/EGU2009-13324-1.pdf, 2009.

Duane, G. S. and Shen, M.-L.: Synchronization of Alternative Models in a Supermodel and the Learning of Critical Behavior, Journal of the Atmospheric Sciences, 80, 1565–1584, 2023.

Duane, G. S. and Tribbia, J. J.: Synchronized chaos in geophysical fluid dynamics, Physical Review Letters, 86, 4298, 2001.

Duane, G. S., Tribbia, J. J., and Weiss, J. B.: Synchronicity in predictive modelling: a new view of data assimilation, Nonlinear Processes in Geophysics, 13, 601–612, https://doi.org/10.5194/npg-13-601-2006, 2006.

Duane, G. S., Wiegerinck, W., Selten, F., Shen, M.-L., and Keenlyside, N.: Supermodeling: Synchronization of alternative dynamical models of a single objective process, Advances in nonlinear geosciences, pp. 101–121, 2018.

Egger, J. and Schilling, H.-D.: On the theory of the long-term variability of the atmosphere, Journal of Atmospheric Sciences, 40, 1073–1085, 1983.

European Centre for Medium-Range Weather Forecasts: ERA5 Reanalysis (0.25 Degree Latitude-Longitude Grid) (Updated monthly) [Dataset]. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory, https://doi.org/https://doi.org/10.5065/BH6N-5N20, 2019.

Frederiksen, J. S.: A unified three-dimensional instability theory of the onset of blocking and cyclogenesis. II. Teleconnection patterns, Journal of Atmospheric Sciences, 40, 2593–2609, 1983.

Gettelman, A. and Morrison, H.: Advanced two-moment bulk microphysics for global models. Part I: Off-line tests and comparison with other schemes, Journal of Climate, 28, 1268–1287, 2015.

Gettelman, A., Liu, X., Ghan, S. J., Morrison, H., Park, S., Conley, A., Klein, S. A., Boyle, J., Mitchell, D., and Li, J.-L.: Global simulations of ice nucleation and ice supersaturation with an improved cloud scheme in the Community Atmosphere Model, Journal of Geophysical Research: Atmospheres, 115, 2010.

Gettelman, A., Bresch, D. N., Chen, C. C., Truesdale, J. E., and Bacmeister, J. T.: Projections of future tropical cyclone damage with a high-resolution global climate model, Climatic Change, 146, 575–585, 2018.

Golaz, J.-C., Larson, V. E., and Cotton, W. R.: A PDF-based model for boundary layer clouds. Part I: Method and model description, Journal of the atmospheric sciences, 59, 3540–3551, 2002.

Gregory, W., Bushuk, M., Adcroft, A., Zhang, Y., and Zanna, L.: Deep learning of systematic sea ice model errors from data assimilation increments, Journal of Advances in Modeling Earth Systems, 15, e2023MS003 757, 2023.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, 2020.

Hoskins, B. J. and Karoly, D. J.: The steady linear response of a spherical atmosphere to thermal and orographic forcing, Journal of the atmospheric sciences, 38, 1179–1196, 1981.

Judt, F.: Insights into atmospheric predictability through global convection-permitting model simulations, Journal of the Atmospheric Sciences, 75, 1477–1497, 2018.

Kang, I.-S.: Influence of zonal mean flow change on stationary wave fluctuations, Journal of Atmospheric Sciences, 47, 141–147, 1990.

Kirtman, B. P. and Shukla, J.: Interactive coupled ensemble: A new coupling strategy for CGCMs, Geophysical research letters, 29, 5–1, 2002.

Lau, N.-C.: Variability of the observed midlatitude storm tracks in relation to low-frequency changes in the circulation pattern, Journal of the atmospheric sciences, 45, 2718–2743, 1988.

Morrison, H. and Gettelman, A.: A new two-moment bulk stratiform cloud microphysics scheme in the Community Atmosphere Model, version 3 (CAM3). Part I: Description and numerical tests, Journal of Climate, 21, 3642–3659, 2008.

Neale, R. B., Chen, C.-C., Gettelman, A., Lauritzen, P. H., Park, S., Williamson, D. L., Conley, A. J., Garcia, R., Kinnison, D., Lamarque, J.-F., et al.: Description of the NCAR community atmosphere model (CAM 5.0), NCAR Tech. Note NCAR/TN-486+ STR, 1, 1–12, 2010.

Palmer, T.: Climate forecasting: Build high-resolution global climate models, Nature, 515, 338–339, 2014.

Palmer, T. and Stevens, B.: The scientific challenge of understanding and estimating climate change, Proceedings of the National Academy of Sciences, 116, 24 390–24 395, 2019.

Park, S. and Bretherton, C. S.: The University of Washington shallow convection and moist turbulence schemes and their impact on climate simulations with the Community Atmosphere Model, Journal of Climate, 22, 3449–3469, 2009.

Pecora, L. M., Carroll, T. L., Johnson, G. A., Mar, D. J., and Heagy, J. F.: Fundamentals of synchronization in chaotic systems, concepts, and applications, Chaos: An Interdisciplinary Journal of Nonlinear Science, 7, 520–543, 1997.

Phillips, A. S., Deser, C., and Fasullo, J.: Evaluating modes of variability in climate models, Eos, Transactions American Geophysical Union, 95, 453–455, 2014.

Raeder, K., Hoar, T. J., Gharamti, M. E., Johnson, B. K., Collins, N., Anderson, J. L., Steward, J., and Coady, M.: A new CAM6+ DART reanalysis with surface forcing from CAM6 to other CESM models, Scientific Reports, 11, 16 384, 2021.

Sardeshmukh, P. D. and Hoskins, B. J.: The generation of global rotational flow by steady idealized tropical divergence, Journal of the Atmospheric Sciences, 45, 1228–1251, 1988.

Schevenhoven, F.: Training of supermodels - in the context of weather and climate forecasting, doctoral thesis, University of Bergen, https://doi.org/https://bora.uib.no/bora-xmlui/handle/11250/2727454, 2021.

Schevenhoven, F., Selten, F., Carrassi, A., and Keenlyside, N.: Improving weather and climate predictions by training of supermodels, Earth System Dynamics, 10, 789–807, 2019.

Schevenhoven, F., Keenlyside, N., Counillon, F., Carrassi, A., Chapman, W. E., Devilliers, M., Gupta, A., Koseki, S., Selten, F., Shen, M.-L., et al.: Supermodeling: improving predictions with an ensemble of interacting models, Bulletin of the American Meteorological Society, 104, E1670–E1686, 2023.

Schevenhoven, F. J. and Selten, F. M.: An efficient training scheme for supermodels, Earth System Dynamics, 8, 429–438, https://doi.org/10.5194/esd-8-429-2017, 2017.

Segura, H., Pedruzo-Bagazgoitia, X., Weiss, P., Müller, S. K., Rackow, T., Lee, J., Dolores-Tesillos, E., Benedict, I., Aengenheyster, M., Aguridan, R., Arduini, G., Baker, A. J., Bao, J., Bastin, S., Baulenas, E., Becker, T., Beyer, S., Bockelmann, H., Brüggemann, N., Brunner, L., Cheedela, S. K., Das, S., Denissen, J., Dragaud, I., Dziekan, P., Ekblom, M., Engels, J. F., Esch, M., Forbes, R., Frauen, C., Freischem, L., García-Maroto, D., Geier, P., Gierz, P., González-Cervera, A., Grayson, K., Griffith, M., Gutjahr, O., Haak, H., Hadade, I., Haslehner, K., ul Hasson, S., Hegewald, J., Kluft, L., Koldunov, A., Koldunov, N., Kölling, T., Koseki, S., Kosukhin, S., Kousal, J., Kuma, P., Kumar, A. U., Li, R., Maury, N., Meindl, M., Milinski, S., Mogensen, K., Niraula, B., Nowak, J., Praturi, D. S., Proske, U., Putrasahan, D., Redler, R., Santuy, D., Sármány, D., Schnur, R., Scholz, P., Sidorenko, D., Spät, D., Sützl, B., Takasuka, D., Tompkins, A., Uribe, A., Valentini, M., Veerman, M., Voigt, A., Warnau, S., Wachsmann, F., Wacławczyk, M., Wedi, N., Wieners, K.-H., Wille, J., Winkler, M., Wu, Y., Ziemen, F., Zimmermann, J., Bender, F. A.-M., Bojovic, D., Bony, S., Bordoni, S., Brehmer, P., Dengler, M., Dutra, E., Faye, S., Fischer, E., van Heerwaarden, C., Hohenegger, C., Järvinen, H., Jochum, M., Jung, T., Jungclaus, J. H., Keenlyside, N. S., Klocke, D., Konow, H., Klose, M., Malinowski, S., Martius, O., Mauritsen, T., Mellado, J. P., Mieslinger, T., Mohino, E., Pawłowska, H., Peters-von Gehlen, K., Sarré, A., Sobhani, P., Stier, P., Tuppi, L., Vidale, P. L., Sandu, I., and Stevens, B.: nextGEMS: entering the era of kilometer-scale Earth system modeling, EGUsphere, 2025, 1–39, https://doi.org/10.5194/egusphere-2025-509, 2025.

Severijns, C. A. and Hazeleger, W.: The efficient global primitive equation climate model SPEEDO V2.0, Geoscientific Model Development, 3, 105–122, https://doi.org/10.5194/gmd-3-105-2010, 2010.

Shen, M.-L., Keenlyside, N., Bhatt, B. C., and Duane, G. S.: Role of atmosphere-ocean interactions in supermodeling the tropical Pacific climate, Chaos: An Interdisciplinary Journal of Nonlinear Science, 27, 2017.

Simmons, A. J., Wallace, J., and Branstator, G. W.: Barotropic wave propagation and instability, and atmospheric teleconnection patterns, Journal of the Atmospheric Sciences, 40, 1363–1392, 1983.

Tegegne, G., Kim, Y.-O., and Lee, J.-K.: Spatiotemporal reliability ensemble averaging of multimodel simulations, Geophysical Research Letters, 46, 12 321–12 330, 2019.

Ting, M. and Lau, N.-C.: A diagnostic and modeling study of the monthly mean wintertime anomalies appearing in a 100-year GCM experiment, Journal of Atmospheric Sciences, 50, 2845–2867, 1993.

van den Berge, L. A., Selten, F. M., Wiegerinck, W., and Duane, G. S.: A multi-model ensemble method that combines imperfect models through learning, Earth System Dynamics, 2, 161–177, https://doi.org/10.5194/esd-2-161-2011, 2011.

Wallace, J. M. and Gutzler, D. S.: Teleconnections in the geopotential height field during the Northern Hemisphere winter, Monthly weather review, 109, 784–812, 1981.

Watt-Meyer, O., Brenowitz, N. D., Clark, S. K., Henn, B., Kwa, A., McGibbon, J., Perkins, W. A., and Bretherton, C. S.: Correcting weather and climate models by machine learning nudged historical simulations, Geophysical Research Letters, 48, e2021GL092 555, 2021.

Watt-Meyer, O., Dresdner, G., McGibbon, J., Clark, S. K., Henn, B., Duncan, J., Brenowitz, N. D., Kashinath, K., Pritchard, M. S., Bonev, B., et al.: ACE: A fast, skillful learned global atmospheric model for climate prediction, arXiv preprint arXiv:2310.02074, 2023.

Weigel, A. P., Knutti, R., Liniger, M. A., and Appenzeller, C.: Risks of model weighting in multimodel climate projections, Journal of Climate, 23, 4175–4191, 2010.

Wiegerinck, W., Burgers, W., and Selten, F.: On the Limit of Large Couplings and Weighted Averaged Dynamics, Understanding Complex Systems, pp. 257–275, https://doi.org/10.1007/978-3-642-33359-0-10, 2013.

Zhang, G. J. and McFarlane, N. A.: Sensitivity of climate simulations to the parameterization of cumulus convection in the Canadian Climate Centre general circulation model, Atmosphere-ocean, 33, 407–446, 1995.