JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2019MS001678

A Reliability Budget Analysis of CESM-DART

Jonathan Eliashiv[1], Aneesh C. Subramanian[2], and Arthur J. Miller[1]

[1]Scripps Institution of Oceanography, ClimaCell, Boulder, CO, USA, [2]Department of Atmospheric and Oceanic Sciences, University of Colorado Boulder, Boulder, CO, USA

Key Points:
- A reliability budget is used to diagnose error in a new prototype coupled ocean-atmosphere reanalysis product, the CESM-DART
- The bias was low, the residual term was large in the budget, indicating incorrect observational error and ensemble variance statistics
- During active MJO periods, the error was lower in regions of suppressed convection, suggesting errors in the convection parameterization

Correspondence to:
A. C. Subramanian,
aneeshcs@colorado.edu

Citation:
Eliashiv, J., Subramanian, A., & Miller, A. J. (2020). A reliability budget analysis of CESM-DART. *Journal of Advances in Modeling Earth Systems*, *12*. e2019MS001678 https://doi.org/10.1029/2019MS001678

Received 22 MAR 2019
Accepted 25 NOV 2019
Accepted article online 28 DEC 2019

Abstract A reliability budget is used to diagnose potential sources of error (departure from observations) in a new prototype coupled ocean-atmosphere ensemble Kalman filter reanalysis product, the Community Earth System Model using the Data Assimilation Research Testbed (CESM-DART). In areas with sufficient observations, the mean bias in zonal wind was generally very low compared to the spread due to ensemble variance, which did not exhibit patterns associated with Northern Hemisphere jet streams but did have regional enhancement over the Maritime Continent. However, the residual term was often the largest contributor to the budget, which is problematic, suggesting improper observational error statistics and inadequately represented ensemble variance statistics. The departure and residual exhibit significant seasonal variability, with a strong peak in boreal winter months, indicating the model's deficiencies during the energetic Northern Hemisphere winter. Ocean temperature contained large error in areas with eddy production indicating inadequate ensemble variance due to poor model resolution. Periods when the Madden-Julian Oscillation (MJO) was active exhibited lower error, especially in the western equatorial Pacific during MJO phases with reduced convection. In contrast, during MJO phases with enhanced convection in that region, the ensemble variance is increased yet the error is comparable to non-MJO conditions, suggesting a controlling effect of the precipitation parameterization. Further studies evaluating the impact of the coupled assimilation procedure on the reliability budget will be illuminating.

Plain Language Summary A reliability budget is used to diagnose error in a new prototype coupled ocean-atmosphere ensemble Kalman filter reanalysis product, the Community Earth System Model using the Data Assimilation Research Testbed (CESM-DART). Although the bias term was low, the residual term was often the largest contributor to the budget, indicating improper observational error statistics and inadequately represented ensemble variance statistics. During periods when the Madden-Julian Oscillation was active, the error was lower, especially in the western equatorial Pacific during Madden-Julian Oscillation phases with reduced convection, suggesting a controlling effect of the convection parameterization.

1. Introduction

Global climate models are well known to contain errors that lead to biases in the mean state and seasonal cycle, as well as to inaccurate representation of variability (e.g., Bretherton et al., 2012; Neale et al., 2013). A great deal of research has been executed to attempt to identify the source of these biases and to alleviate them by developing better parameterizations or increasing resolution (e.g., Gent et al., 2010; Hwang & Frierson, 2013; Jochum, 2009; Neale et al., 2008; Phillips et al., 2004; Subramanian & Zhang, 2014; Zhang & McFarlane, 1995). But the problem remains and is still being attacked so that new approaches to find and fix these errors are actively being sought because of their critical importance to improving weather and climate forecasting.

A promising technique for potentially unravelling some of the sources of model error is the reliability budget of Rodwell et al. (2016). In this strategy, the differences that arise between a model's ensemble members and the target observations of forecasts are broken down into the parts due to mean bias, ensemble spread, observational error, and residuals. This can be applied for any selected field in the model state vector, for any time period, and for any region. The results can potentially shed light on what processes are most likely to be responsible for error growth, where and when they arise, and how they relate to changing background states associated with climate modes of variability (Rodwell et al., 2018).

Here we use the reliability budget to study errors in a new prototype coupled reanalysis product, the Community Earth System Model (CESM) in the Data Assimilation Research Testbed (DART) (Karspeck et al., 2018). CESM-DART invokes a weakly coupled data assimilation scheme using an ensemble adjustment Kalman filter (EAKF) (Anderson et al., 2009). Applying the reliability budget to this ensemble-based coupled reanalysis could provide new insights into the climate processes that conspire to generate model errors in CESM. This is because the surface fluxes, which are the primary means of communication between the atmosphere and ocean, are computed as part of the assimilation scheme itself (Penny & Hamill, 2017). And the estimates of these fluxes from coupled climate models are often cited as containing the biases and random errors that result in inaccurate representations of climate processes and internal modes of variability (e.g., Arnold et al., 2015).

Our reliability budget analysis takes a careful look at how the CESM-DART performs overall during the sequential updating that occurs every 6 hr over the available time period of 1971–1982. What regions contain the most mean bias development? Where does ensemble spread due to instability dominate the model-data mismatch? How does the importance of different terms in the reliability budget change with the seasons or from year to year? How does the model perform in data-rich versus data-sparse regions? We then focus on model performance in representing the Madden-Julian Oscillation (MJO), which is an important and predictable mode of climate variability known to be often poorly represented in climate models (e.g., Kim et al., 2009; Subramanian et al., 2011; Waliser et al., 2009). We break up the budget into times when MJO is active compared to when it is not active. Does the presence of MJO increase or reduce error? Which phases of MJO generate the most error? These and other issues are addressed in this study, as we try to relate this statistical breakdown of errors to focal points in the dynamics that require attention for future model improvements. This will also allow us to potentially identify where we can trust the CESM-DART analysis the most for specified regions, seasons, or climate mode activities.

In the next section, we present the basics of the reliability budget (including an expanded version of the budget that is supported by a simple model diagnosis in the appendix), explain the CESM-DART reanalysis in more details, and describe how MJO conditions are isolated in the model. In section 3, the results are presented for applying the budget to atmospheric and oceanic variables, as well as for MJO versus non-MJO conditions. Section 4 contains a summary discussion and conclusion.

2. Methodology

2.1. Reliability Budget

The framework for our analysis is the reliability budget of Rodwell et al. (2016) that can partition the mean square differences of the ensemble mean error (i.e., relative to observed data values) into bias, ensemble variance, and observational uncertainty contributions, along with a residual that indicates a breakdown of the assumptions used in deriving the budget. This is applied to an ensemble of "forecasts," which in this case are the ensemble members used to compute the ensemble-mean reanalysis fields of CESM-DART, at each update time. Averaging the budget terms over time and/or space gives an aggregate perspective of the model performance and the source of mismatch between forecast and observed.

Under the assumptions of perfect reliability and a perfect knowledge of the observation-error characteristics (Rodwell et al., 2016), the reliability budget can be written as follows:

$$\overbrace{\overline{(m-o)^2}}^{\text{Departure}} = \overbrace{\overline{m-o}^2}^{\text{Bias}} + \overbrace{\overline{e_m^2}}^{\text{Ensemble variance}} + \overbrace{\overline{e_o^2}}^{\text{Observation uncertainty}} + \underbrace{(Var(o) - \underline{e_o^2}) + (Var(m) - \underline{e_m^2}) - 2Cov(m,o)}_{\text{Residual}} \tag{1}$$

where o is an observation (including error), m is the ensemble mean model state variable, $\overline{e_m^2}$ is the ensemble variance of the model system, and $\overline{e_o^2}$ is the variance of the observational uncertainty. The overbar indicates averaging over all observations for all times. Variance and covariance for arbitrary variables, x and y, are defined as

$$Var(x) = \overline{x^2} - \bar{x}^2$$
$$Cov(x,y) = \overline{xy} - \bar{x}\,\bar{y}$$

so that one can see that equation (1) simply corresponds to the expansion of the variance of differences between the model-ensemble-mean and observations (i.e., the departure from observations) as follows:

$$\begin{aligned} Var(m-o) &= Var(m) + Var(o) - 2Cov(m,o) \\ \overline{(m-o)^2} &= \overline{m-o}^2 + Var(m) + Var(o) - 2Cov(m,o). \end{aligned} \tag{2}$$

If the reliability budget does not balance so that the residual is nonzero, then we have either incorrectly assigned observational error statistics or we have inadequately represented forecast variance statistics, or both. We shall see below that the residual is indeed frequently nonzero, and significant, in our analysis of CESM-DART.

As a further exploration of the "structural bias" in the model (Williamson et al., 2015), we expand the model ensemble mean state variable, m, into its true value in nature, T, plus two error terms, one representing the mean error, termed the structural bias, $f(x)$, for any state variable x, plus the random error around that mean, e_m. The observation data can additionally be broken down into T plus observational uncertainty e_o, which is generally prescribed in the development of the data assimilation scheme. Using these expansions, derivations for the structural bias can be written as follows:

$$m \equiv T - e_m + f(x) \tag{3}$$

$$o \equiv T - e_o \tag{4}$$

$$\begin{aligned} Var(m) &= Var(T) + Var(e_m) + Var(f(x) - 2Cov(T, e_m)) \\ &\quad + 2Cov(T, f(x)) - 2Cov(e_m, f(x)) \\ &= Var(T) + Var(e_m) + Var(f(x)) + 2Cov(T, f(x)) \\ Var(o) &= Var(T) + Var(e_o) - 2Cov(T, e_o) \\ &= Var(T) + Var(e_o) \\ Cov(m,o) &= Cov(T + f(x), T) \\ &= Var(T) + Cov(f(x), T) \\ \delta &\equiv m - o \\ Var(\delta) &= Var(e_m) + Var(e_o) + Var(f(x)) \\ Var(e_o) &= Var(\delta) - Var(m) + Cov(m,o) + Cov(T, f(x)) \end{aligned}$$

If the structural bias is assumed to be linear or approximately linear, the strength of the bias and ensemble variance can be isolated and estimated:

$$\begin{aligned} f(x) &\approx Ax + B \\ Cov(o,x) &= Cov(T,x) \\ Cov(m,x) &\approx Cov(T,x) + AVar(x) \\ A &\approx \frac{Cov(\delta,x)}{Var(x)} \end{aligned}$$

$$Var(e_o) \approx Var(\delta) - Var(m) + Cov(m,o) + ACov(o,x) \tag{5}$$

$$Var(e_m) \approx Var(\delta) - Var(e_o) - A^2 Var(x) \tag{6}$$

Both (5) and (6) can then be directly calculated from the observations and the model reanalysis products. The results can then be used to breakdown their influence on the residual, that is, determine possible reasons for the unbalanced budget.

We will apply the relations from equations (5) and (6) to CESM-DART in the next section. The purpose of using the above technique is to test whether the true ensemble variance is of the same order as the variance produced by the model output, as well as whether the observational uncertainty is specified correctly. However, the approximations invoked may not hold true, so we apply the technique to a simplified theoretical

model (Lorenz, 1963) where we can specify the errors as known values in both model and observations and thereby test under what conditions the linearity approximation for structural bias is valid. This analysis is presented in the appendix. The results reveal that one of the limitations found for this approximation to the budget terms is that model error variance, e_m^2, and the observational uncertainty variance, e_o^2, need to be of the same order of magnitude. This analysis does show that using this approach allows for a quantification of linear bias, A, with less than 8% error when using at least 1,000 data points and less than 20% error in the quantification of background bias, B, when using at least 40,000 points, even when sampled randomly. This quantification may not be as effective when using the CESM-DART product of the 1970s because there is a limitation due to the low observational density. However, for systems using modern sensors and data, the structural bias can likely be identified effectively using this linearized estimate.

2.2. CESM-DART

The DART system is an open-source community developed software used to implement the EAKF data assimilation technique into different models (Anderson et al., 2009). The application of DART to the CESM (Karspeck et al., 2018) provides a novel depiction of climate data for the 12-year period that it was run and made available to us (1971–1982). CESM-DART uses CESM1.1.1, which is a coupled model, which combines the Community Atmospheric Model Version 4 (Gent et al., 2011), with the Community Land Model (CLM4), the Community Ice Code (CICE4), CESM River Transport Model, and the LANL Parallel Ocean Program (POP2) for a coupled earth climate model (Hurrell et al., 2013). CESM-DART is run with a 30-member ensemble, with data assimilated every 6 hr for atmospheric fields and 24 hr for oceanic components. The components are coupled together every 24 hr. ACARS and BUFR data (see Karspeck et al., 2018, for the data sources) is used to adjust CAM and POP state vectors. Aircraft, radiosonde, and a few satellite measurements of wind and temperature are used for the atmosphere model. Bottle, DBT, STD, CTD, MBT, and XBT measurements of temperature and salinity are used for the ocean model. These reanalysis fields data are compared to the individual assimilated observations before the next Kalman filter update is applied at each sequential update time and for each observation.

2.3. MJO Index

We use MJO events as one measure of intraseasonal coupled climate variability. To isolate MJO events, the velocity potential MJO (VPM) index is computed using the methods described by Ventrice et al. (2013). Velocity potential is calculated at pressure level 200 hPa, which serves as smoother measure of convective activity. Anomalies of the velocity potential and zonal wind anomalies at pressure levels 200 and 850 hPa are then band-pass filtered between 20 and 100 days and symmetrically averaged between 15°N and 15°S. The combined empirical orthogonal function (Wheeler & Hendon, 2004) is then calculated for the resulting fields, which provides the structures of the coherent meridional variability. The sum of squares of the two leading principal components provide the VPM index, which when exceeding 1.5 is used to indicate the occurrence of MJO events.

3. Results

3.1. Reliability Budget

Since the reliability budget is applied in observation space, we first examine where the analysis can be effectively used during the available CESM-DART reanalysis period, 1971–1982. Figure 1 shows the observation density of all the different variable types assimilated over 1971–1982 in CESM-DART. Each variable uses multiple products from ACARS and BUFR for atmosphere and ocean measurements, respectively. Observational density is poor in the Southern Hemisphere but reasonably high in many areas of the Northern Hemisphere and in the tropical oceanic regions.

The reliability budget was initially applied to each of the different model state variables, which were first mapped to a regular 1° ×1 ° grid on which the observations were binned, then averaged vertically and plotted globally. As an example of these results, which had many large-scale features in common among the model state variables, Figure 2 (left) shows the terms of the budget for zonal winds. The top plot shows the root-mean-square error (departure from observations) for the entire 12-year reanalysis. Notice that this can only be plotted where there were data to be assimilated (plotted at the bottom as observational density), so some remote areas are blank. The two largest terms that contribute to the error are ensemble variance and the residual. Surprisingly, the ensemble variance does not exhibit patterns associated with Northern Hemisphere jet streams, but it does have regional enhancement over the Maritime Continent region where deep

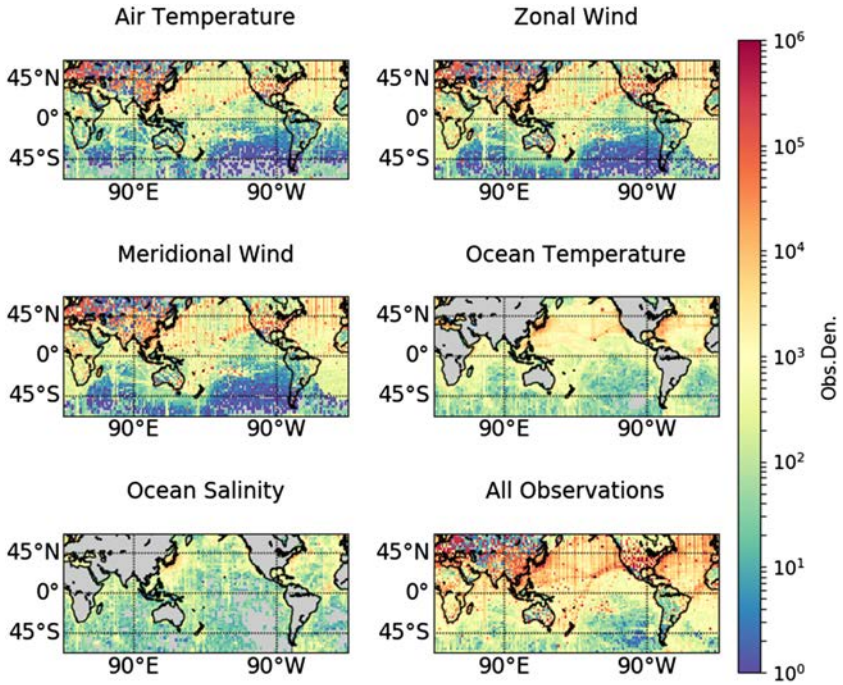

Figure 1. Observation density of atmospheric (temperature, zonal wind, and meridional wind) and oceanic (temperature and salinity) measurements, right to left and top to bottom, that are assimilated into CESM using DART. All observations together are plotted in the bottom right panel.

convection frequently occurs. The bias is low in most areas, although significant bias can be found in some regions. The uncertainty in the input observations (which is specified before assimilation) is generally very small.

In the initial representation of the reliability budget (Figure 2, left), large amounts of departure can be found in two different types of regions. The first and most dramatic are locations on the southern edges of where data is available in the middle-to-high-latitude Southern Hemisphere. There are energetic flows in those regions but very limited observations to constrain the atmospheric activity. This will be further explored below. Another location of higher departure is in the Northern Pacific and Atlantic basin. This region is not associated with an increase in ensemble variance or observational uncertainty nor is the increased departure visible in the bias. This indicates that those regions display a higher variability in the model or observation noise than can be explained by either the ensemble variance or the observational uncertainty. For those areas, we will use the expanded budget analysis derived from equations (5) and (6) in order to isolate their response.

3.2. Expanded Reliability Budget

Figure 2 (right) shows the same reliability budget but including the terms $Var(e_m)$, $Var(e_o)$, and $\sum_i A_i^2 Var(x_i)$, where subscript i indicates the spatial location, from equations (6) and (5). In Figure 2 (right), e_o is derived from data under the assumption of linear bias. Though the estimates for e_o and e_m are not likely to be quantitatively accurate, as found in the appendix for a simplified model, the spatial structures of the variability of e_o and e_m are likely to be meaningful. However, the plots show that e_m is not spatially similar to the ensemble variance, suggesting that the model does not produce enough variability within the ensemble. The structure of e_o resembles the residual because the observational uncertainty and ensemble variance cannot span the entire variability of the climate system. The variability of $\sum_i A_i^2 Var(x_i)$ shows that there is small amount of structural bias distributed through the system, but it is found to be negligible in comparison to the departure. It is therefore nonlinear and not readily explored in this linearized framework.

This initial foray into exploring the reliability budget motivates us to pursue a more detailed examination of the budget by considering different regions, timescales, variables, and processes, in order to attempt to isolate dominant mechanisms controlling the error development. In particular, we focus later on the MJO as a target

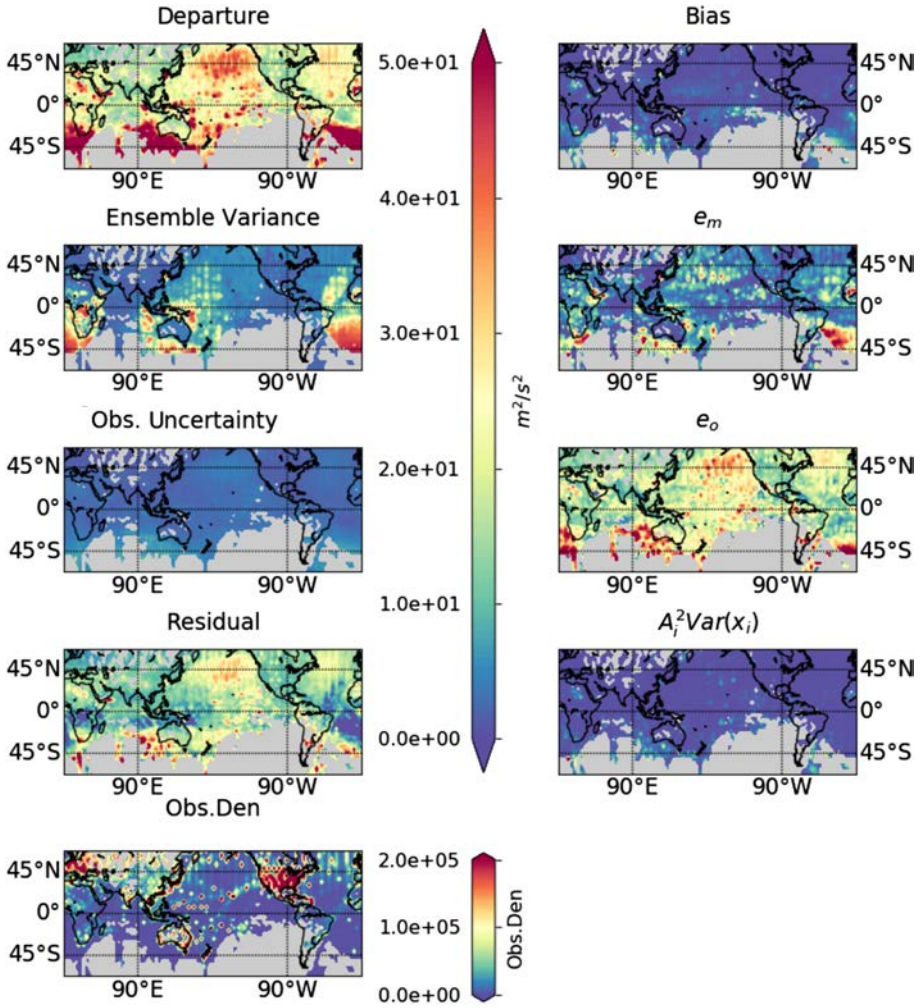

Figure 2. (left) Reliability budget terms for atmospheric zonal velocity, vertically averaged. From top to bottom, Departure (total error variance), bias of the ensemble mean, ensemble variance, observational uncertainty, residual. observation density is also plotted for reference. (right) Expanded reliability budget terms for model error variance, observational error variance, and linearly estimated Bias.

climate process, since it gives insight into the growth of errors associated with atmospheric convectively coupled equatorial waves on intraseasonal timescales, which are key to extended-range predictions.

3.3. Single Dimension Exploration

Figure 3 shows equation (1) for zonal wind as a function of longitude, latitude, height, and time. In the figure, the histogram in the background shows the observation density for the measurements in each dimension. The main feature that is evident is a reduction of the entire budget with higher observation density when plotted versus latitude or longitude.

The correlation between observation density and each term of the reliability budget is shown in Table 1. In the table r_x^2 represents the R^2 value calculated against $log(n + 1)$, where n is observation density in the x domain. We find that departure and observational uncertainty have the strongest dependence on observation density when considered as a function of latitude. There is also some weak longitudinal relationship between departure and observational density but essentially no vertical or temporal relationship.

Figure 3 shows two interesting features in the time domain. There is a general trend toward lower departures in the latter half of the record as the observational density increases. There is also a clear seasonal cycle in the departure and residual, likely indicating a seasonal cycle in the ability of the model to produce observed levels of variability in the Northern Hemisphere where the data are concentrated. This is further explored in Figure 4, which shows the seasonal cycle climatology (after subtracting the annual mean) and the anomalies

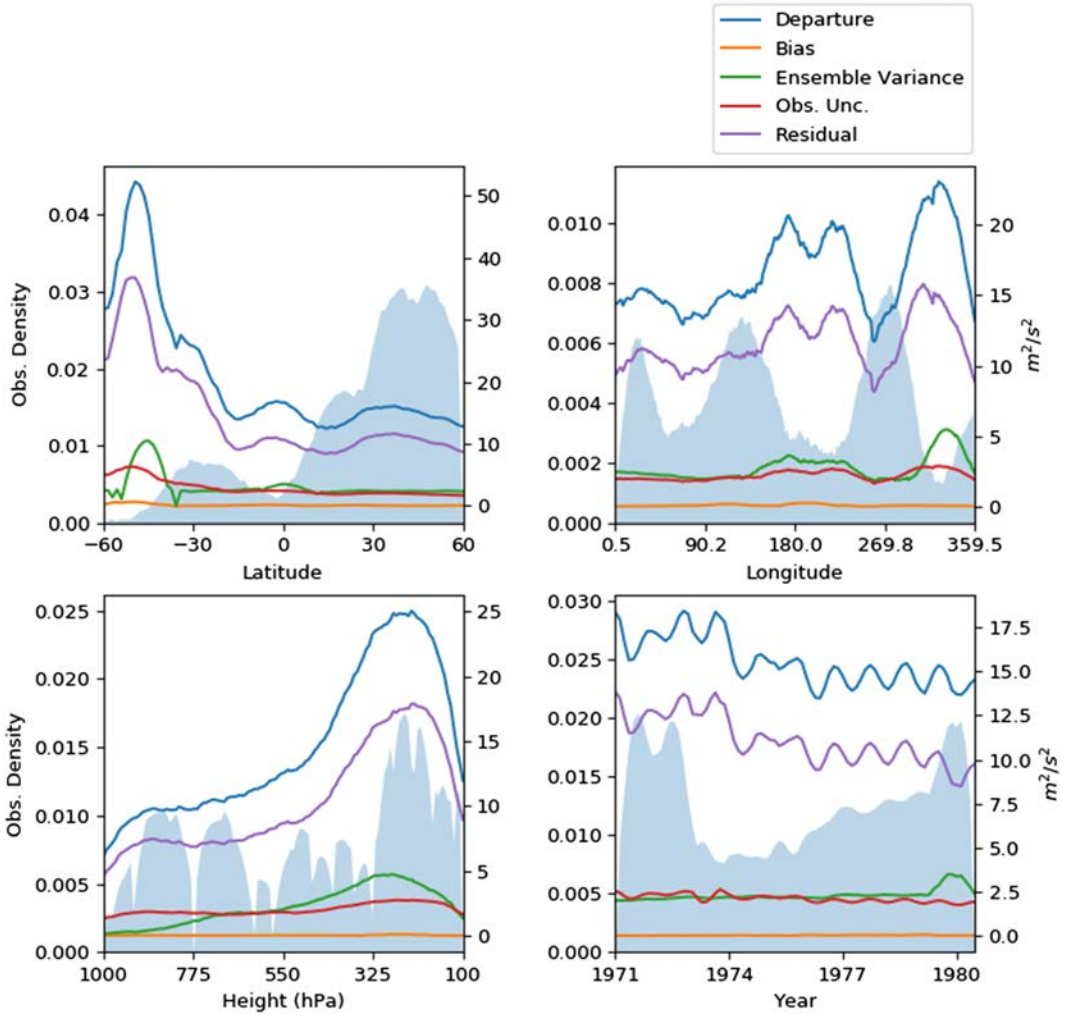

Figure 3. Reliability budget for atmospheric zonal velocity, plotted as a function of latitude, longitude, time, and height, and otherwise averaged in the other dimensions. Shaded regions represent the observation density at each grid point.

(after removing the seasonal cycle and annual mean) of the budget. The departure and residual both have the most seasonal variability, with a strong peak in winter months, and weaker signals in other months, indicating the model's deficiencies during the energetic Northern Hemisphere winter. The anomalies of the budget, however, do not seem to correlate with observational density in time or with any particular climate mode of variability during this 12-year period. Note that the combination of seasonal cycle and interannual variability account for less than 15% of the total budget.

Table 1
Squared Correlation Between Terms in the Reliability Budget (From Figure 4) and $\log(n+1)$ *in the Variable Specified by Each Row*

	Departure	Bias	Ens. Var.	Obs. Unc.	Residual
r^2_{lat}	0.56	0.30	0.17	0.69	0.45
r^2_{lon}	0.19	0.10	0.36	0.14	0.09
r^2_{height}	0.05	0.01	0.02	0.07	0.05
r^2_{time}	0.03	0.00	0.15	0.05	0.06

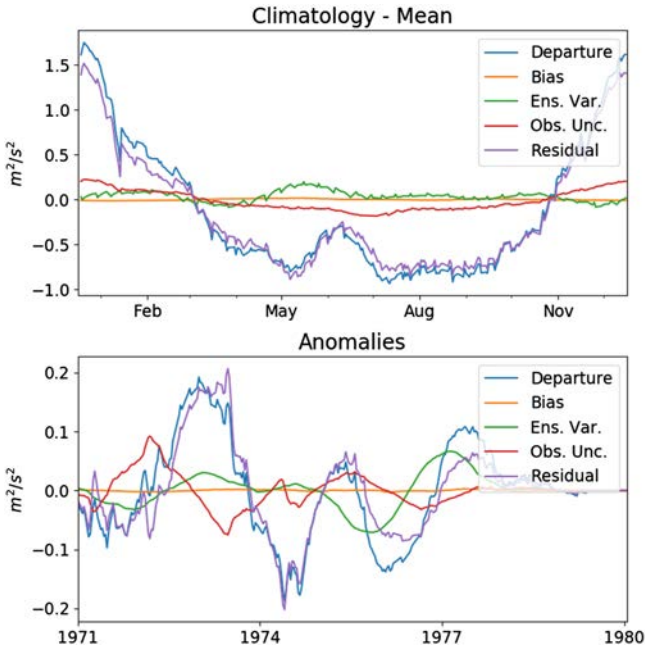

Figure 4. Reliability budget for atmospheric zonal velocity plotted as (top) a seasonal cycle and (bottom) anomalies around the seasonal cycle, after subtracting out the annual mean, and otherwise averaged over latitude, longitude, and height.

To quantify the affect of observation density, first-order approximations of its effect are derived as follows:

$$\begin{aligned} n &\equiv n(\mathbf{x}) = n(t,\theta,\phi,z) \\ R(n) &= R\left(n(\mathbf{x})\right) \\ \frac{\partial R}{\partial n} &= \sum_i \frac{\frac{\partial R}{\partial x_i}}{\frac{dn}{dx_i}} \\ \overline{\frac{\partial R}{\partial n}} &\equiv \int \frac{\mathrm{d}n}{\mathrm{d}\mathbf{x}} \frac{\partial R}{\partial n} \mathbf{dx} \Big/ \int \frac{\mathrm{d}n}{\mathrm{d}\mathbf{x}} \mathbf{dx} \end{aligned} \tag{7}$$

where n is observation density, t is time, θ is longitude, ϕ is latitude, z is height, and R is the respective reliability budget term.

Using this derivation, Figure 5 shows the relative contributions of observation density, time, and spatial dimensions. It can easily be deduced that observation density is by far the dominant producer of error for most aspects of the reliability budget. It also confirms that the overwhelming contribution of the strong meridional dependence on the budget was from observation density, and not meridional biases. Surprisingly, Figure 5 also shows that departure has a very strong dependence on height overwhelmingly over other variables. Though the bias terms are dominated mostly by observation density, their contributions to the budget are 2 orders of magnitude lower than any other contribution and are therefore negligible in this context.

Both ensemble variance and observational uncertainty rise with height, but surprisingly, ensemble variance increases with observation density as well. These effects may be caused by sudden influx of observations in areas that were not well covered, which would lead to a sudden model correction. In order to explore the mechanism for this rise, the model would have to be examined with data input in more regular intervals in both time and space for comparison.

3.4. Ocean Temperature Exploration

Figure 6 shows the reliability budget for ocean temperatures taken around the globe and throughout the water column, although these are concentrated in the upper ocean. There is a large production of error (departure) in areas with eddy production in western boundary currents, as well as in equatorial regions where there are tropical instability waves. These errors are reflected in the ensemble variance, and to a lesser degree in bias, but have very small amounts of observational uncertainty. In fact, the departure and ensemble variance are highly correlated to each other, showing that areas with unstable currents are being regularly corrected to observations, as would be expected.

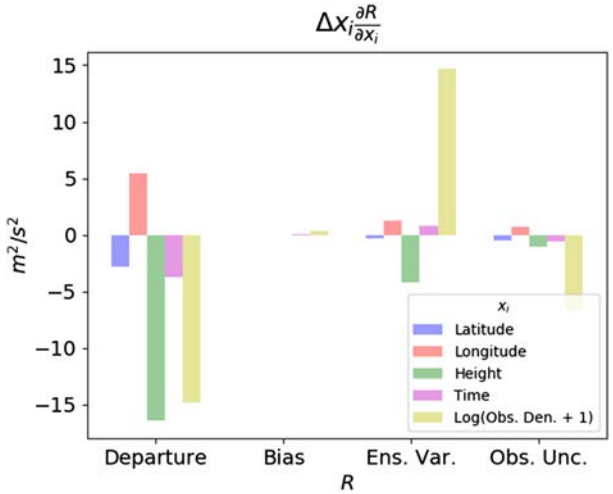

Figure 5. Sensitivity of the terms in the reliability budget to latitude, longitude, height, time, and observation density.

The areas with unstable current features have approximately the same order of magnitude of residual and ensemble variance. This shows that ensemble variance is actually underestimated through the dynamical processes of the model, indicative of the coarse resolution being incapable of producing energetic eddies. Additionally, the assimilated measurements are given heavy weight even in the unstable regions and thereby prevent the growth of error. In more quiescent regions, such as in the subtropical gyres, the departures are small.

Table 2 looks at the effect of ocean water temperature on the reliability budget. Areas with warmer waters have larger amounts of errors than areas with colder waters. This is found in departure, bias, and ensemble variance. The ensemble variance surpasses the departure for the warmest waters. This shows that the model produces exaggerated variability in especially warm waters, but weaker variability otherwise.

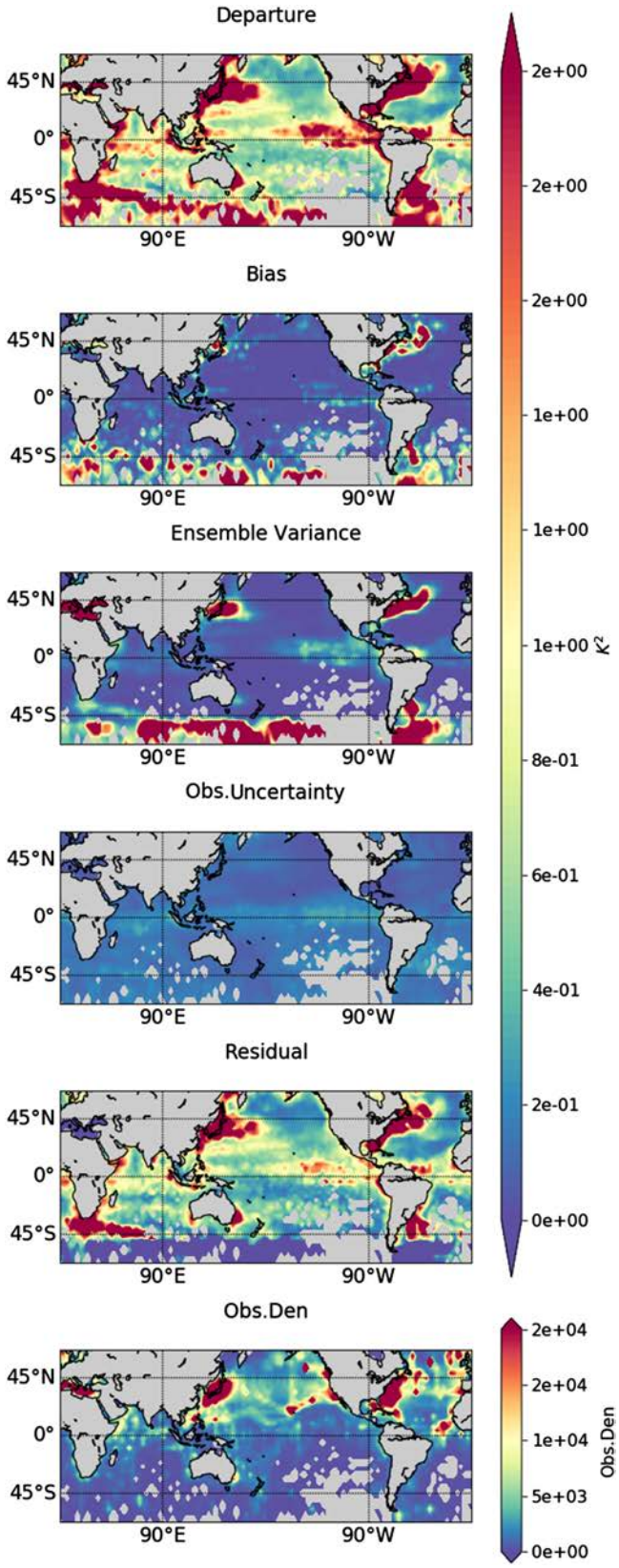

Figure 6. As in Figure 2 but for ocean temperature, vertically averaged.

Table 2
Reliability Budget Terms for Ocean Temperatures Averaged Globally for the Indicated Range of Ocean Temperature

	$\theta < 10$ K	10 K$\leq \theta <$ 20 K	20 K$\leq \theta$
Departure	1.69	2.18	2.70
Bias	0.08	0.17	0.19
Ens. Var	0.09	1.50	4.00
Obs.Unc	0.11	0.10	0.09
Residual	0.64	0.04	−1.59

3.5. Intraseasonal Exploration

In a related study using CESM-DART (Eliashiv et al., 2018), the data assimilation in CESM significantly improved the performance of MJO events, which suggested that the flux coupling at the ocean-atmosphere interface may be important in organizing the convective activity inherent in MJO. Focusing on MJO events, which have well-known mechanisms involved, may give us added insight into the processes leading to the error growth in CESM-DART. Using the VPM index, time periods when MJO are active (VPN > 1.5) are separated from those when it is not. We then averaged the reliability budget terms over the tropical (30°S to 30°N) Indian (40–100°E) and western Pacific Ocean (130–170°E) regions, where MJO is most convectively active. Ocean temperature is included for this analysis since MJO events convectively couple the oceanic and atmospheric systems within the tropical Indian and Pacific basins.

Table 3 shows that there is, surprisingly, a slight reduction in error (departure) for both atmospheric zonal winds and ocean temperature during MJO events, even though the atmosphere can be more convectively vigorous at these times. To further explore this reduction in departure during MJO, the VPM index was compared to the anomalies of the global zonal wind reliability budget. They were found to be uncorrelated, indicating that this reduction in error in the Indian and Pacific basins is not simply a coincidental effect that occurs circumglobally during times of MJO.

The composite MJO structure and phasing is frequently shown as spatial maps of velocity and convective activity during eight phases as it propagates around the globe. We next broke up the reliability budget into each of the eight phases of MJO, compared to times when MJO is not active. Figure 7 shows how the different MJO phases influence the reliability budget for the Indian and Pacific basins, specifically within 850 ± 50 and 200 ± 50 hPa. These heights are chosen as they are heights at which divergence and convergence specific to the MJO structure occur. There are two key characteristics to notice for this figure. The first is that upper tropospheric errors are reduced for these two basins for all phases, with or without MJO events, in contrast to the full column results shown in Table 3 as well as in Figures 3 and 5, which show an increase in error with height. The second feature is that the departure exhibits its lowest values for MJO Phases 2 to 3 in the lower troposphere of the western Pacific. These are the times of least active convection locally there, as the MJO propagates from the Indian Ocean toward the western Pacific Ocean.

Table 3
Reliability Budget Terms for Zonal Winds and Ocean Temperatures During MJO and Non-MJO Conditions Averaged Over Western Equatorial Pacific Ocean and Indian Ocean

		Departure	Bias	Ens.Var.	Unc.	Residual
Pacific (no MJO)	U (m^2/s^2)	17.	0.037	0.26	2.3	13.
	θ (K^2)	2.0	0.013	0.003	2.3	−0.25
Pacific (MJO)	U (m^2/s^2)	16.	0.047	0.26	2.2	13.
	θ (K^2)	1.9	0.011	0.003	2.2	−0.29
Indian (no MJO)	U (m^2/s^2)	14.	0.023	0.25	1.9	12.
	θ (K^2)	1.6	0.057	0.004	6.0	−4.5
Indian (MJO)	U (m^2/s^2)	14.	0.032	0.25	1.8	12.
	θ (K^2)	1.4	0.017	0.004	4.3	−2.9

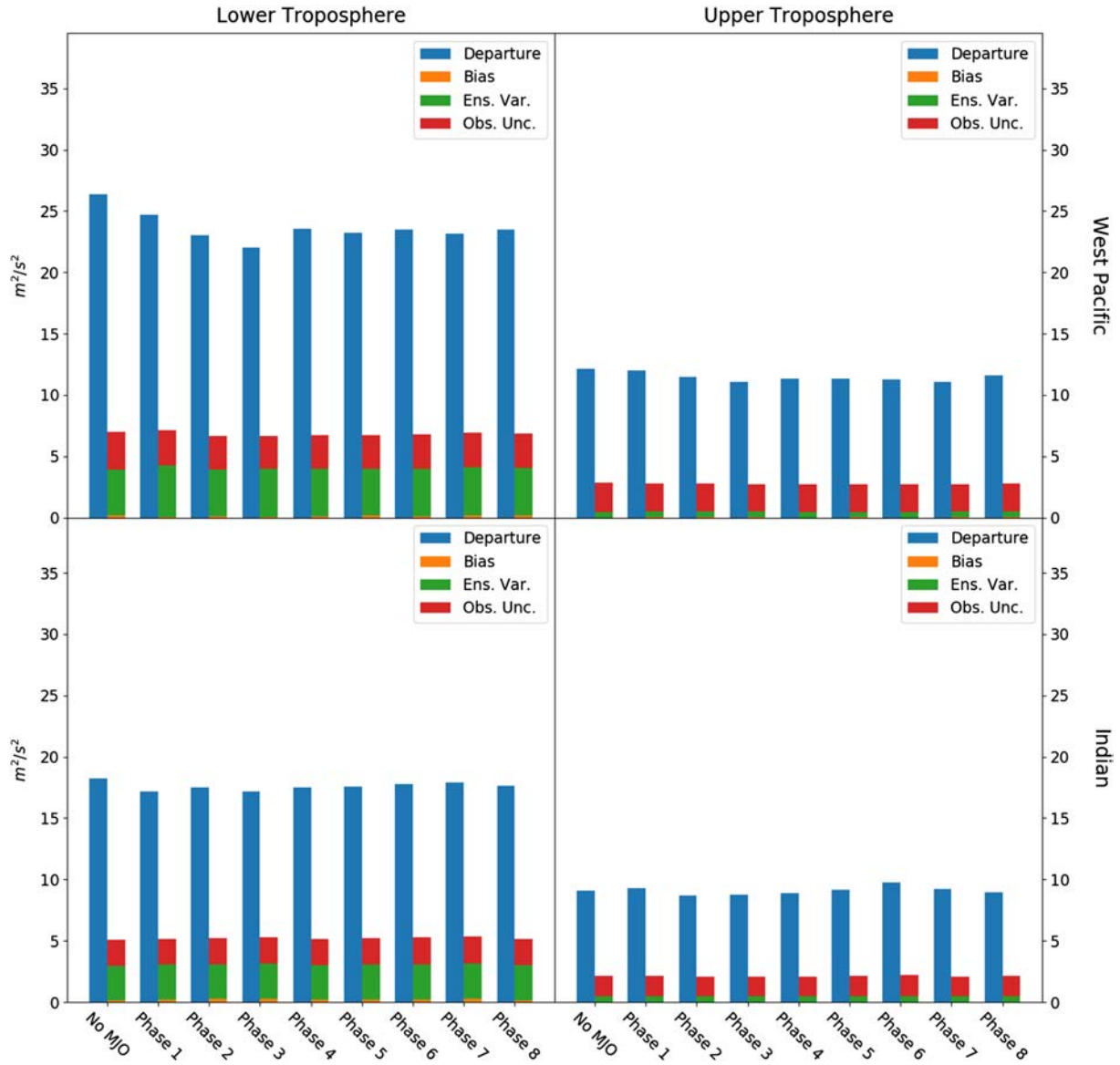

Figure 7. Reliability budget terms for each phase of the MJO and for non-MJO conditions, averaged for the western Pacific and Indian basins for lower tropospheric (850 ± 50 hPa) and upper tropospheric (200 ± 50 hPa) zonal winds.

As a final exploration for the mechanism that causes the reduction in error, Figure 8 shows how zonal velocity affects the occurrence of error in the western Pacific basin. Plotted as histograms are the available zonal wind observations throughout the troposphere in the equatorial western Pacific region for non-MJO conditions as well for each MJO phase. The zonal wind distribution peaks at negative values due to the easterly trade winds averaged over this area for all MJO phases as well as for non-MJO conditions. Also plotted are the terms in the reliability budget for the zonal winds that land in the histogram bins, averaged over all available observations in the vertical in that region, with the full budget plotted for non-MJO conditions and the differences from the non-MJO budget plotted for each phase. The top figure shows that without MJO events, higher departures occur when the zonal velocity magnitude increases, with easterlies producing more error than westerlies. It additionally shows that there is no strong dependence of the error production on wind speed except for exceedingly strong winds, for which there are not many measurements to support the accuracy of the budget.

An important feature during the MJO phases is that the departure is reduced for easterly winds during Phases 1, 2, 3, 7, and 8. These are times when the MJO convective activity in the region is reduced. This

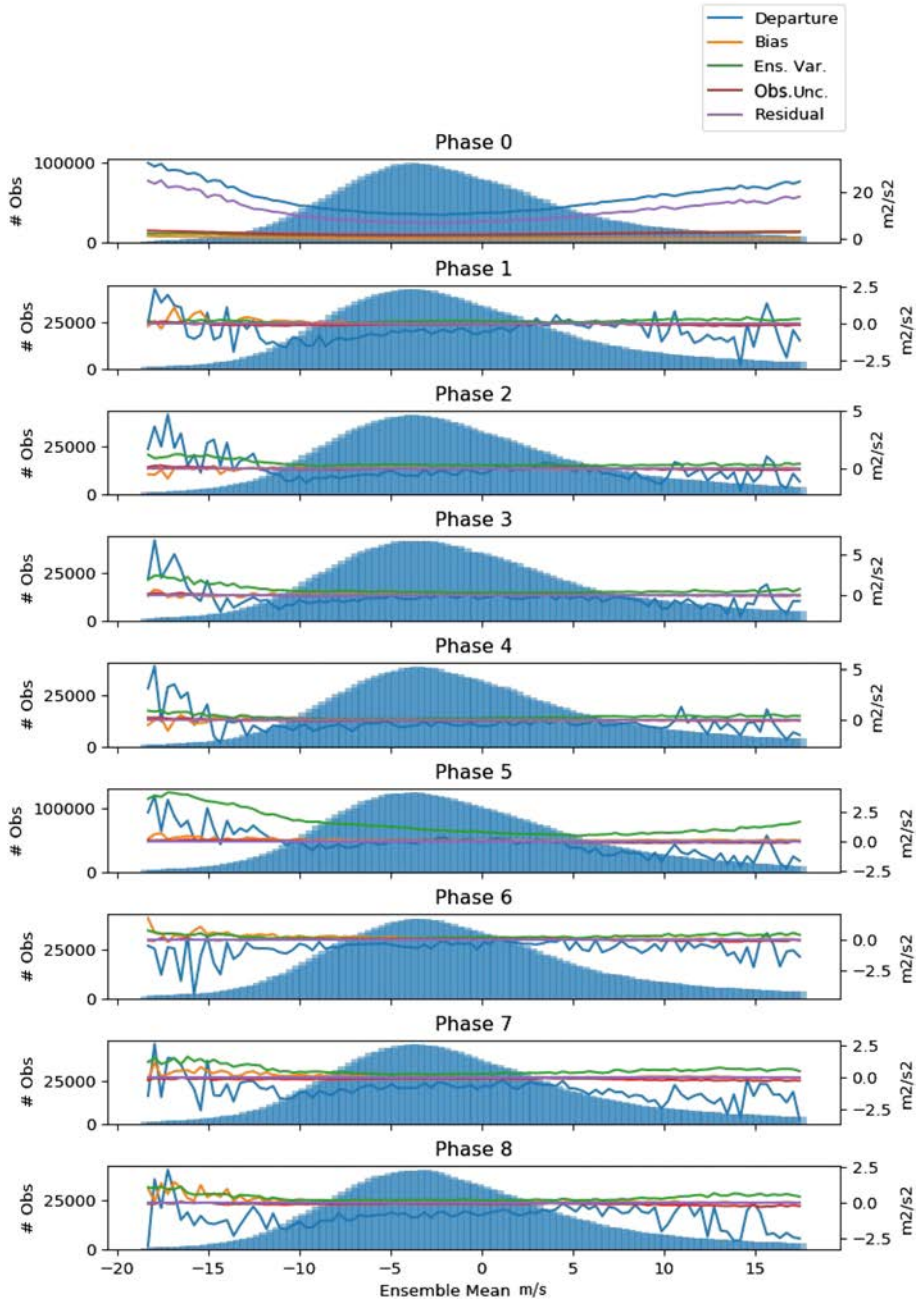

Figure 8. Histograms of zonal winds (blue shading), for all available observations in the vertical in the western equatorial Pacific region, for (top panel) non-MJO conditions and (lower panels) each phase of the MJO. Also plotted as colored lines are the corresponding reliability budget terms for the zonal winds that occur in each bin of the histogram, to show the impact of wind strength on the budget. The top panel is the full budget for non-MJO time conditions, and remaining panels are plotted as differences from the top panel for each phase of MJO.

indicates that the model representation of organized MJO propagation is sufficiently good that it is able to reduce errors in the reanalysis during MJO events. Surprisingly, the ensemble variance is not concomitantly reduced, indicating that it is not simply the nonlinearity in the local model dynamics that is controlling the reduction in these departures. In other words, lower convection does not equate to both better fit to observations and lower ensemble spread in the model at the same times in this region. Note, in contrast, that the ensemble variance does increase significantly in Phase 5, when MJO convective activity is maximum in this region, yet the departure is not changed from non-MJO (normal) conditions. This suggests that the

model precipitation parameterization may be responsible for producing the ensemble spread through feedbacks with the dynamics, thereby limiting the overall skill of the reanalysis in this region for both MJO and non-MJO conditions.

4. Summary and Conclusion

We used a reliability budget (Rodwell et al., 2016) analysis to diagnose potential sources of error (termed departure) in the ensemble mean from an ensemble Kalman filter-based CESM-DART climate reanalysis of the time period 1971–1982 (Karspeck et al., 2018). We focused our study on identifying where we can trust the CESM-DART analysis the most for specified regions, seasons, or climate mode activity. We used zonal winds as a typical example of our results for other atmospheric variables. In areas where there were sufficient numbers of observations, the mean error bias was generally very low compared to the spread due to ensemble variance, indicating the important controls exerted by the unstable dynamics in the atmospheric flows. The ensemble variance did not exhibit patterns associated with Northern Hemisphere jet streams but did have regional enhancement over the Maritime Continent. However, the largest ensemble variance was in the middle-to-high-latitude Southern Hemisphere, where there are energetic flows with limited observations to control them. The residual terms were often the largest contributors to the budget, which is problematic. This suggests that either the statistics of observational uncertainty, which was very small and specified in the application of EAKF, was improperly assigned, or the model inadequately represented ensemble variance statistics, or both. Clearly, increasing observational uncertainty, which corresponds in this case to targeting more proper recognition of model representation error, can help improve the application of the EAKF to CESM. This may entail using nonconstant, process-specific values of observational error tailored to the spatial region where key processes are not properly represented in CESM, such as for convection in the West Pacific Warm Pool and the Maritime Continent. Additional experimental designs are needed to better assess these shortcomings in the model, especially for recent time periods where more data are available.

We also attempted to develop a strategy to explore the relative importance of model error and observational uncertainty in contributing to the residual. We used an "identical twin" framework with a simple, but nonlinear, model (Lorenz, 1963) to prescribe known errors for both model and observations, such that the observations contain only random error and the model contains both random error and structural bias (assumed to be linear). However, we found that the estimate of structural bias will only be sufficiently accurate when the observational and model errors are of similar magnitude, which is generally not applicable in more complicated models. Applying the strategy qualitatively to CESM-DART suggested that the model does not produce enough variability within the ensemble in many regions around the globe. A larger ensemble in the DART formalism may help to alleviate this issue. Additionally, since CESM-DART includes only multiplicative inflation of ensemble covariance, adding a stochastic contribution to the inflation procedure may help to generate a more realistic ensemble spread.

Note that the bias in our reliability budget is distinct from the mean bias that develops in a free-running climate model in statistical equilibrium after nonlinear interactions have acted to balance the flows. In our analysis, the bias corresponds to the recurring systematic drift of the ensemble of "forecasts" away from the "observed" reanalysis initial state and toward the targeted observation, in a tangent-linear sense. There may be no relation between the spatial patterns seem in our bias maps and the climatologist mean bias maps from a long-term current-climate simulation using CESM. But patterns in the reliability budget bias (and the other terms) may reveal the tendencies that suggest why the model consistently behaves incorrectly in some region, for a particular variable, during a certain season, or while a particular climate mode of variability is active. Targeting the physical mechanisms in those regions may help to improve the mean bias development in CESM.

Toward this goal, we next examined the temporal and spatial variability of the reliability budget. The departure and residual exhibit significant seasonal variability, with a strong peak in winter months, and weaker signals in other months, indicating the model's deficiencies during the energetic Northern Hemisphere winter. The year-to-year variability in the budget exhibited no obvious relationship to any climate mode. A sensitivity analysis revealed that, over this time period with limited data, the observation density predominantly controlled the quality of the reliability budget, which explained some of features seen in the latitudinal and longitudinal structures of the budget, since far more data were available over land areas of the Northern Hemisphere.

We also studied the ocean temperature field and, not surprisingly, found a large production of error (departure) in areas with eddy production in western boundary currents, as well as in equatorial regions where there are tropical instability waves. These areas with unstable current features have approximately the same magnitude of residual and ensemble variance. This is consistent with the coarse resolution of the model being incapable of dynamically producing energetic eddies thereby underestimating ensemble variance.

We then studied intraseasonal variability by identifying periods when MJO was active and compared them to nonactive time periods to provide insight into the growth of errors associated with atmospheric convectively coupled equatorial waves. We found a slight reduction in error (departure) for both atmospheric zonal winds and ocean temperature during MJO events, even though this includes times when the atmosphere is more convectively vigorous than normal. After breaking down the budget by the phase of MJO, we noted that the departures exhibit their lowest values in the lower troposphere of the tropical western Pacific for MJO Phases 2 to 3, which are the times of least active convection there. This suggests that weaker convective activity in the atmosphere allows a locally more skillful model representation of the flows. We also found that in the tropical western Pacific during the more convectively quiescent phases of MJO (1, 2, 3, 7, and 8) the departure is reduced for moderate-to-strong easterly wind conditions. In contrast, during Phase 5, when convection is enhanced by MJO in that region, the ensemble variance is increased over non-MJO times, but the skill is left unchanged. Taken together, this suggests that the model precipitation physics produces enhanced ensemble spread through feedbacks with the dynamics, thereby limiting the overall skill of the reanalysis in this region for both MJO and non-MJO conditions.

These results provide a novel perspective of the quantitative performance of CESM-DART in providing a reliable estimate of the atmospheric and oceanic state during 1971–1982, for specified regions, seasons, and climate mode activity. But further studies evaluating the impact of the coupled ensemble-based assimilation procedure on the reliability budget are needed. For instance, since the study was executed in observations space, there was no assessment of the reliability of the surface fluxes because they were not included in the data assimilation procedure (cf. Eliashiv et al., 2018 who found enhanced MJO activity may be due to the surface flux coupling in CESM-DART). Additionally, evaluations using ensemble-based reanalysis products with more properly specified observational uncertainty and the high-density observations that are available in more recent decades may also provide clearer perspectives on how to pinpoint the processes in the model that need the most improvement in their parameterization.

Appendix A: Error Analysis in Lorenz-63

We use an "identical twin" framework with a simple, but nonlinear, model to explore the relative importance of model error to observational uncertainty. We apply the Lorenz (1963) system, which is derived from the dominant modes of the solution to Rayleigh-Benard problem which describes convective tropospheric behavior as follows:

$$\phi(x,z,t) \propto x_1(t)\sin(\pi x/k)\sin(\pi z/l)$$
$$\theta(x,z,t) \propto \sqrt{2}x_2(t)\cos(\pi x/k)\sin(\pi z/l) - x_3(t)\sin(2\pi z/l)$$
$$\frac{\mathrm{d}}{\mathrm{d}t}x_1 = \sigma(x_2 - x_1)$$
$$\frac{\mathrm{d}}{\mathrm{d}t}x_2 = x_1(\rho - x_3) - x_2$$
$$\frac{\mathrm{d}}{\mathrm{d}t}x_3 = x_1x_2 - \beta x_3$$

where ϕ is the stream function, θ is temperature anomalies, $\sigma = 10$ is the Prandtl number, $\rho = 28$ is the rescaled Rayleigh number, and $\beta = 8/3$ is derived from the wavenumbers k and l. Using this system, data are derived from 50 high-resolution cycles (loops around the entire attractor) made using a forward Runga-Kutta fourth-order time step. The state variable x_1 is then ascribed as truth and used as input to create both an erroneous model state and and uncertain observation.

Errors are then prescribed and added to both model and observations following equations (3) and (4), so that the observations contain only random error, e_o, and the model contains both random error, e_m, and structural bias, At (linear growth with time) + B (constant background bias). Unless otherwise specified, the initial default values of e_o, e_m, A, and B are set to be $\sqrt{var(x_1)}/10$, where $\sqrt{var(x_1)}$ is the root-mean-square of the anomalies produced by the model. The random error is generated as a uniform distribution.

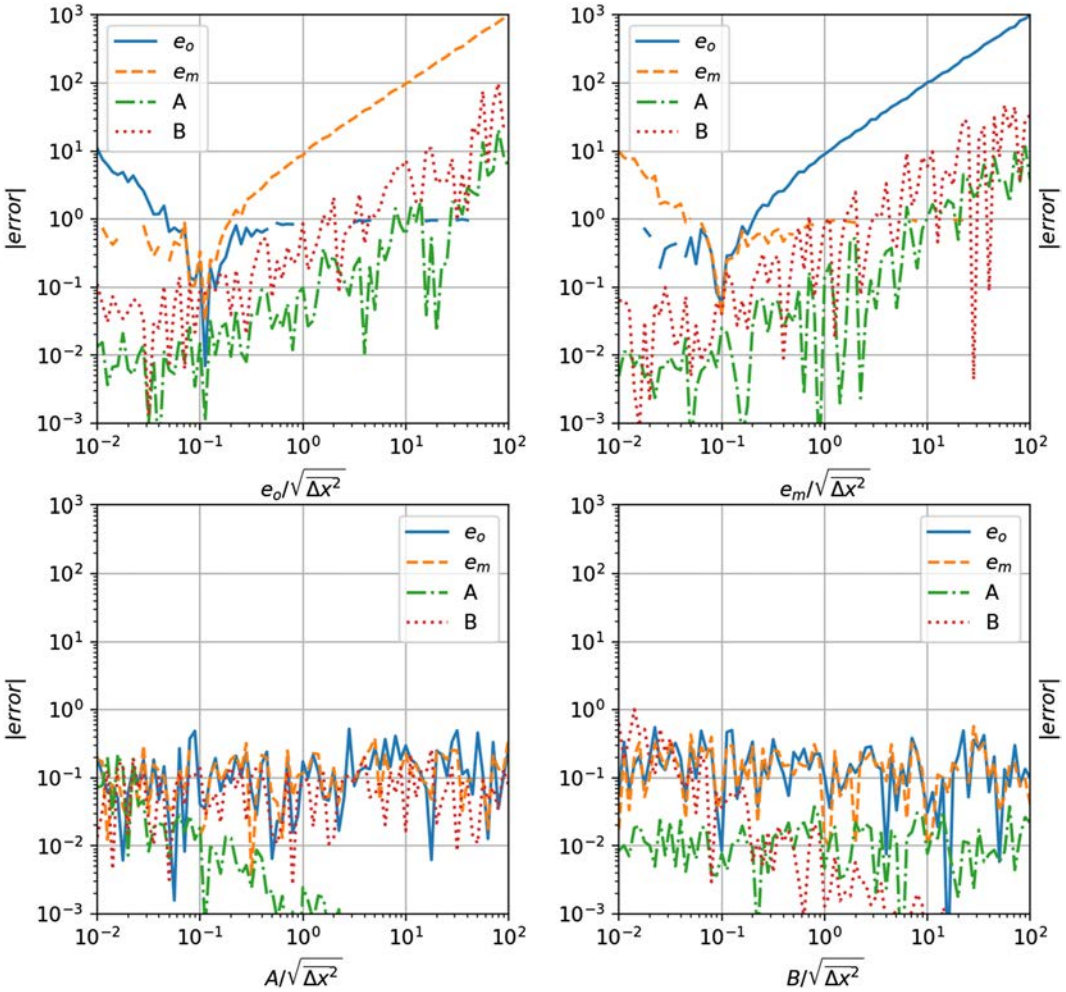

Figure A1. Error in the estimates of e_o, e_m, A, and B as a function of (a) e_o, (b) e_m, (c) linear bias, A and (d) background bias, B. In each subplot, values for the other parameters are set to $\sqrt{\overline{x^2}}/10$. These plots are made using 1,000 randomly selected values of truth.

Using equations (5) and (6), we can estimate values of e_o, e_m, A, and B, and examine whether or not they are close to their prescribed values. Figure A1 shows the error in the estimation of these values while varying the magnitude of the specified errors. For example, in the upper left plot, if the observational error, e_o, is very low, the estimate of A and B are both very good (lower than 0.1), but the estimates of e_o and e_m are large and unusable. But if the observational error is near 0.1, then all the variables are estimated well. If the observational error, e_o, is comparable to the variability of the system (1%.0), the estimate of A is still useful (lower than 0.1), but the estimates of B, e_o, and e_m are large and inaccurate. Likewise, in the upper right panel, the same situation occurs for e_m, but with the roles of e_m and e_o reversed.

The upper panels of Figure A1 clearly show that if e_o or e_m increase to be as large as the natural variance in the system, the error in estimating the other variables increases to levels that render the estimates invalid. Interesting, however, when A or B grows in magnitude, the other value estimates are hardly affected, while the bias estimates themselves grow more and more accurate.

Another major point found by inspecting Figure A1 is that when e_o and e_m are of similar magnitude, the estimation becomes the most accurate for all four parameters. To isolate this effect, Figure A2 shows the structure of the error when varying both e_o and e_m. The plot shows a minimum in error when e_o and e_m are equal in magnitude and increasing error when their magnitudes begin to differ. This shows that in order for the approximations in equations (5) and (6) to hold true, e_m and e_o must be approximately the same magnitude.

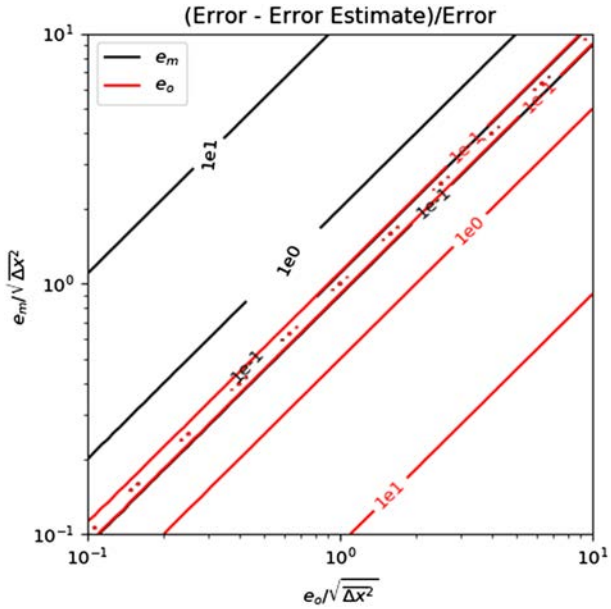

Figure A2. Error in estimating e_o (red) and e_m (black) as a function of both e_o and e_m. Values for the other parameters are set to $\sqrt{x^2}/10$.

These plots were made using 1,000 randomly selected values of truth. We repeated this experiment using regularly spaced sampling time intervals, and with e_m derived using the known e_o, all which showed no major effects on the experiment result. We also explored how the number of observations affect our approximate estimate of the structural bias. Figure A3 shows the effect of sampling density on error estimation. This plot was made using randomly generated values for e_o, e_m, A, and B (within 20% of the canonical value of 0.1). For each plotted point, the experiment was run 100 times, and the average error in estimation was recorded. The figure shows that even with O(10^6) observations, e_o and e_m will have estimation errors that range from from 10% to over 100% of the value itself. Exceptions to this are of course found in Figure A2, when e_o and e_m are nearly equal, as discussed above. Figure A3 additionally shows that structural biases, A, can be found

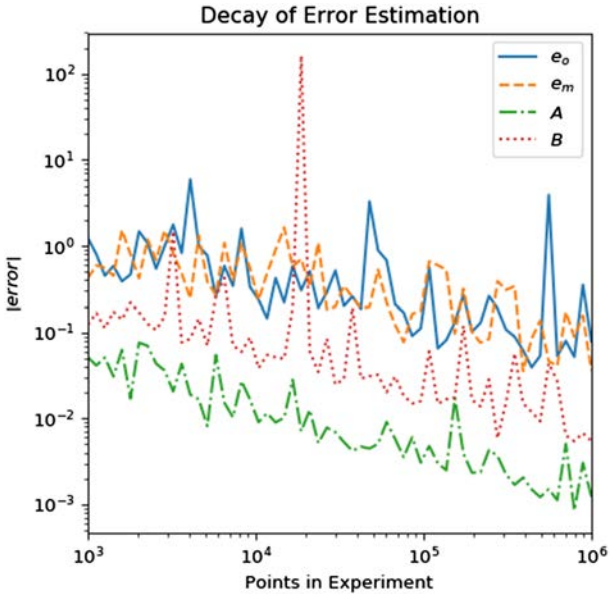

Figure A3. Error in estimating e_o, e_m, A, and B with increasing data points (observations). For each plotted point, simulations are created with random values of the above parameters 100 times, and estimates from each simulation are averaged, as described in the text.

with less than 8% estimation error even with only O(10^3) randomly sampled observations, or more. Finally, background biases, *B*, can be estimated with less than 10% error when using at least O(10^5) points.

Glossary

ACARS	Aircraft Communications Addressing and Reporting System
BUFR	Binary Universal Form for the Representation
CESM	Community Earth Systems Model
DART	Data Assimilation Research Testbed
EAKF	Ensemble adjustment Kalman filter
hPa	hectopascal
MJO	Madden-Julien Oscillation
VPM	Velocity potential MJO index
WMO	World Meteorological Organization

Acknowledgments
This study forms a portion of the Ph.D. dissertation of J. A. This research was only possible because of the immense effort by Alicia Karspeck and the NCAR team, who generously provided us with the CESM-DART data set and access to their observational database. This research was supported by a grant from NOAA Climate Variability and Prediction Program (NA14OAR4310276) and the NSF Earth System Modeling Program (OCE1419306). High-performance computing support from Yellowstone (ark:/85065/d7wd3xhc) was provided by NCAR's Computational and Information Systems Laboratory (CISL), which is sponsored by the National Science Foundation. The model output and observational data were created, assembled, and studied by Karspeck et al. (2018, QJRMS). These CESM-DART model output and observational data are available on the archive servers of the Data Support Section of CISL at NCAR. We thank the two anonymous referees who provided perceptive and insightful comments that significantly improved the clarity and interpretation of results in our manuscript.

References

Anderson, J., Hoar, T., Raeder, K., Liu, H., Collins, N., Torn, R., & Avellano, A. (2009). The data assimilation research testbed: A community facility. *Bulletin of the American Meteorological Society*, *90*(9), 1283–1296.

Arnold, N. P., Branson, M., Kuang, Z., Randall, D. A., & Tziperman, E. (2015). Mjo intensification with warming in the superparameterized CESM. *Journal of Climate*, *28*(7), 2706–2724.

Bretherton, C., Balaji, V, Delworth, T, Dickinson, R., Edmonds, J., Famiglietti, J., & Smarr, L. (2012). *A national strategy for advancing climate modeling*. Washington, DC: The National Academies Press.

Eliashiv, J., Subramanian, A. C., & Miller, A. J. (2018). Tropical climate variability in the Community Earth System Model: Data Assimilation Research Testbed. *Climate Dynamics*, *54*(1–2), 793–806. https://doi.org/10.1007/s00382-019-05030-6

Gent, P. R., Danabasoglu, G., Donner, L. J., Holland, M. M., Hunke, E. C., Jayne, S. R., et al. (2011). The community climate system model version 4. *Journal of Climate*, *24*(19), 4973–4991.

Gent, P. R., Yeager, S. G., Neale, R. B., Levis, S., & Bailey, D. A. (2010). Improvements in a half degree atmosphere/land version of the CCSM. *Climate Dynamics*, *34*(6), 819–833.

Hurrell, J. W., Holland, M. M., Gent, P. R., Ghan, S., Kay, J. E., Kushner, P. J., et al. (2013). The community Earth system model: A framework for collaborative research. *Bulletin of the American Meteorological Society*, *94*(9), 1339–1360.

Hwang, Y.-T., & Frierson, D. M. (2013). Link between the double-intertropical convergence zone problem and cloud biases over the Southern Ocean. *Proceedings of the National Academy of Sciences*, *110*(13), 4935–4940.

Jochum, M. (2009). Impact of latitudinal variations in vertical diffusivity on climate simulations. *Journal of Geophysical Research*, *114*, C01010. https://doi.org/10.1029/2008JC005030

Karspeck, A. R., Danabasoglu, G., Anderson, J., Karol, S., Collins, N., Vertenstein, M., et al. (2018). A global coupled ensemble data assimilation system using the community Earth system model and the data assimilation research testbed. *Quarterly Journal of the Royal Meteorological Society*, *144*(717), 2404–2430.

Kim, D, Sperber, K, Stern, W, Waliser, D, Kang, I-S, Maloney, E, et al. (2009). Application of MJO simulation diagnostics to climate models. *Journal of Climate*, *22*(23), 6413–6436.

Lorenz, E. N. (1963). The mechanics of vacillation. *Journal of the Atmospheric Sciences*, *20*(5), 448–465.

Neale, R. B., Richter, J. H., & Jochum, M. (2008). The impact of convection on ENSO: From a delayed oscillator to a series of events. *Journal of climate*, *21*(22), 5904–5924.

Neale, R. B., Richter, J., Park, S., Lauritzen, P. H., Vavrus, S. J., Rasch, P. J., & Zhang, M. (2013). The mean climate of the community atmosphere model (CAM4) in forced SST and fully coupled experiments. *Journal of Climate*, *26*(14), 5150–5168.

Penny, S. G., & Hamill, T. M. (2017). Coupled data assimilation for integrated Earth system analysis and prediction. *Bulletin of the American Meteorological Society*, *97*(7), ES169–ES172.

Phillips, T. J., Potter, G. L., Williamson, D. L., Cederwall, R. T., Boyle, J. S., Fiorino, M., et al. (2004). Evaluating parameterizations in general circulation models: Climate simulation meets weather prediction. *Bulletin of the American Meteorological Society*, *85*(12), 1903–1916.

Rodwell, M., Lang, S., Ingleby, N., Bormann, N, Hólm, E, Rabier, F, et al. (2016). Reliability in ensemble data assimilation. *Quarterly Journal of the Royal Meteorological Society*, *142*(694), 443–454.

Rodwell, M. J., Richardson, D. S., Parsons, D. B., & Wernli, H. (2018). Flow-dependent reliability: A path to more skillful ensemble forecasts. *Bulletin of the American Meteorological Society*, *99*(5), 1015–1026.

Subramanian, A. C., Jochum, M., Miller, A. J., Murtugudde, R., Neale, R. B., & Waliser, D. E. (2011). The Madden–Julian Oscillation in CCSM4. *Journal of Climate*, *24*(24), 6261–6282.

Subramanian, A. C., & Zhang, G. J. (2014). Diagnosing MJO hindcast biases in NCAR CAM3 using nudging during the dynamo field campaign. *Journal of Geophysical Research: Atmospheres*, *119*(12), 7231–7253.

Ventrice, M. J., Wheeler, M. C., Hendon, H. H., Schreck, C. J. III, Thorncroft, C. D., & Kiladis, G. N. (2013). A modified multivariate Madden–Julian Oscillation index using velocity potential. *Monthly Weather Review*, *141*(12), 4197–4210.

Waliser, D., Sperber, K., Hendon, H., Kim, D., Maloney, E., Wheeler, M., et al. (2009). MJO simulation diagnostics. *Journal of Climate*, *22*(11), 3006–3030.

Wheeler, M. C., & Hendon, H. H. (2004). An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Monthly Weather Review*, *132*(8), 1917–1932.

Williamson, D., Blaker, A. T., Hampton, C., & Salter, J. (2015). Identifying and removing structural biases in climate models with history matching. *Climate dynamics*, *45*(5-6), 1299–1324.

Zhang, G. J., & McFarlane, N. A. (1995). Role of convective scale momentum transport in climate simulation. *Journal of Geophysical Research*, *100*(D1), 1417–1426.