Tendency Bias Correction in Coupled and Uncoupled Global Climate Models with a Focus on Impacts over North America

Y. Chang[12], S. D. Schubert[13], R. D. Koster[1], A. M. Molod[1] and H. Wang[3]

[1]Global Modeling and Assimilation Office, NASA/GSFC, Greenbelt, MD

[2]Goddard Earth Sciences Technology and Research, Morgan State University, MD

[3] Science Systems and Applications, Inc., Lanham, MD

Accepted 20 November 2018 in J. Climate

Corresponding author: Siegfried D. Schubert, siegfried.d.schubert@nasa.gov

Abstract

We revisit the bias correction problem in current climate models, taking advantage of state-of-the-art atmospheric reanalysis data and new data assimilation tools that simplify the estimation of short-term (6-hourly) atmospheric tendency errors. The focus is on the extent to which correcting biases in atmospheric tendencies improves the model's climatology, variability, and ultimately forecast skill at subseasonal and seasonal time scales. Results are presented for the NASA/GMAO GEOS model in both uncoupled (atmosphere only) and coupled (atmosphere-ocean) modes.

For the uncoupled model, the focus is on correcting a stunted North Pacific jet and a dry bias over the central US during boreal summer – long-standing errors that are indeed common to many current AGCMs. The results show that the tendency bias correction (TBC) eliminates the jet bias and substantially increases the precipitation over the Great Plains. These changes are accompanied by much improved (increased) storm track activity throughout the northern middle latitudes. For the coupled model, the atmospheric TBCs produce substantial improvements in the simulated mean climate and its variability, including a much reduced SST warm bias, more realistic ENSO-related SST variability and teleconnections, and much improved subtropical jets and related sub-monthly transient wave activity.

Despite these improvements, the improvement in subseasonal and seasonal forecast skill over North America is only modest at best. The reasons for this, which are presumably relevant to any forecast system, involve the competing influences of predictability loss with time and the time it takes for climate drift to first have a significant impact on forecast skill.

1. Introduction

Substantial progress has been made over the last few decades to improve the ability of climate models to reproduce the observed climate. For example, Flato et al. (2013) provide an overview of the quality of the CMIP5 climate models (IPCC, 2013), including a synthesis of our confidence in the ability of models to simulate various features of the 20th century climate including means, various modes of variability, trends, and extremes. They conclude that overall, climate models are indeed getting better in simulating climate (e.g., compared to CMIP3 models), providing greater confidence in the appropriateness of these models for climate change studies.

Nevertheless, despite these overall improvements, current climate models are far from perfect, and specific biases appear to be especially detrimental to forecast skill on subseasonal to seasonal time scales, our focus here. For example, during boreal summer, the middle latitude jets serve as wave-guides for Rossby waves entering North America and Europe (e.g., Schubert et al. 2011, Wang et al. 2017). Any deficiencies in the simulation of the summer jets would therefore likely affect our ability to predict Rossby wave impacts on weather and climate extremes over the northern continents. During boreal winter, the hydroclimate of North America is strongly affected by moisture influx from the North Pacific (e.g., Wang and Schubert 2014) that is linked to North Pacific synoptic systems steered by the jet stream. Indeed, the occurrence of drought along the west coast of the U.S. is especially sensitive to the strength and position of the planetary waves, especially the west coast ridge (e.g., Seager et al. 2015); such waves are linked to modes of internal atmospheric variability such as the Pacific North American (PNA) pattern as well as to ENSO and other tropical SST anomalies (e.g., Seager et al. 2015; Seager and Henderson 2016.

The degree of verisimilitude required in simulating these modes (as well as the mean state) for improving forecast skill at subseasonal and seasonal scales is unclear. Corrections to model biases can be made “after the fact”; operational forecasts can be post-processed to deal with climate drift estimates determined from long histories of reforecasts (e.g., Kirtman et al. 2014), and biases in variability can be corrected through such methods as quantile mapping (e.g., Cannon, 2016. Such approaches, however, can only go so far – they cannot correct, for example, for the complete absence of a critical atmospheric mode or linkage during a forecast. Indeed, certain forecast deficiencies can only be avoided by improving the accuracy of the model simulation itself.

Given the difficulty of addressing certain model biases quickly through model improvement, some have considered a stopgap approach: introducing empirically determined “on-line” corrections to the model’s tendency equations. A number of studies have examined the impact of such statistical corrections to early operational and/or simplified numerical models with a focus on developing methods for improving weather forecasts (e.g., Leith 1978; Schemm and Faller 1986; Saha 1992; DelSole and Hou 1999). In a recent study, Danforth et al. (2007) addressed the problem of estimating and correcting model errors using two simplified but realistic GCMs. They found that online state-independent corrections result in significant improvements in the skill of weather forecasts, improvements that are larger than those obtained with *a posteriori* corrections. They further found that state-dependent corrections resulted in worse prediction skill due to sampling errors in the estimation of the full covariance matrix, though they were able to obtain some improvements by localizing the covariance matrix, or alternatively by introducing an SVD-based formulation of the correction operator. We note that another approach, based on historical analogs, that takes into account the possible state-dependence of errors has been shown to be successful (when applied after the fact) in reducing biases in the planetary-scale waves in medium range forecasts (Yu et al. 2014a,b).

In this study, we revisit the bias correction problem, employing a state-of-the-art reanalysis (MERRA-2) and modern data assimilation tools to correct the systematic model tendency errors in both uncoupled (atmospheric general circulation model, AGCM) and coupled (atmospheric-ocean general circulation model, AOGCM) versions of the NASA Global Modeling and Assimilation Office (GMAO) GEOS model. Rather than weather forecasting, our focus here is on examining the extent to which correcting the short-term model tendency biases leads to improvements in some of the GEOS model's long-standing mean climate biases (e.g., in the North Pacific Summer Jet (NPSJ), the boreal winter stationary waves, and the Intertropical Convergence Zone (ITCZ)) – biases that are indeed found in a number of AGCMs and AOGCMs. In addition, we examine whether there are any associated improvements in the simulation of weather and climate variability as well as in the forecast skill attained over North America at subseasonal to seasonal time scales.

Section 2 describes the methodology used, the GEOS model, and the experiments performed. Section 3a (Section 3b) shows the impact of the bias corrections on the climatic means, variances, and covariances simulated in the uncoupled (coupled) versions of the model, and Section 3c examines their impact on subseasonal and seasonal forecast skill over North America. Discussion and conclusions are provided in Section 4.

2. Methodology and Model Experiments

a. Estimating the tendency biases

The GEOS data assimilation system currently uses an increment analysis update (IAU) procedure designed to reduce analysis-induced initial shocks in the model forecast phase of the assimilation cycle (Bloom et al. 1996). The IAU procedure incorporates a constant analysis increment due to each atmospheric analysis, gradually (over the course of the analysis period) as a forcing term in the model tendency equations. Any non-zero long-term average of the IAU increments is what we define here as the “tendency bias” of the model – a bias that presumably causes the model to drift away from the reanalysis climate during the course of a long-term forecast. To be clear on terminology, our use of the word “bias” refers to time mean differences between the model forecasts and observations (or reanalysis) *that are functions of forecast lead-time*. As such, the tendency bias (as defined above) and the model’s climatological bias (that obtained from a free-running climate simulation) represent the two end points of the bias evolution (also referred to here as the drift), with the former measuring how the model initially starts to drift away from the observed climate, and the latter measuring where it ends up (after the model loses all memory of the initial state). A key question we address here is to what extent does correcting the initial bias correct the climatological bias of the free running model.

The IAU approach can be applied “after the fact” by using an existing reanalysis and a sequence of short term forecasts to estimate the increments, correcting the model accordingly at each time step – basically mimicking the IAU procedure used during an assimilation. Such an approach, referred to as “replay” (Orbe et al. 2017; Takacs et al. 2018), can be used with an existing reanalysis to force a model to remain close to that reanalysis at each time step. The tendency bias correction (hereafter, TBC) method is essentially a replay but instead of applying the increment from a specific forecast-analysis difference, applies a long-term averaged increment (retaining the diurnal and annual cycles) at every time step. Details of the methodology are provided in Appendix A. In this way, TBC takes advantage of an existing assimilation or previously generated replay to estimate the long-term mean model tendency biases and uses them as additional forcing terms in the model equations. It is assumed that the TBCs reflect error growth that is linear and therefore should provide a reasonable estimate of the biases in the model tendencies, subject to any observational/analysis biases (e.g., Xue et al., 2013; Bhargava et al. 2018).

Since the uncoupled model used here is the same as that used to produce MERRA-2 (though run at a lower resolution), the tendency bias terms for u, v, T, q, and p_s are taken directly from the MERRA-2 increments for the period 1980-2015, averaged to the lower resolution (nominally 1^o). This, while likely not optimal as compared to using a new replay to MERRA-2 at the lower resolution, was done for practical reasons. Recent work suggests some dependence of the results on resolution (e.g. Achuthavarier et al. 2017).

Our initial attempt at correcting the coupled model was to simply correct the atmospheric fields (i.e., to impose the tendency bias terms derived from MERRA-2) and then couple the corrected atmosphere to the ocean. This however resulted in spurious feedbacks to the corrections in the tropics that apparently result from a mismatch between the atmospheric biases in the coupled and uncoupled models. We instead found it necessary to carry out a replay to MERRA-2 while running in coupled mode. It is important to note that, even for the coupled model, we correct only atmospheric quantities[1]. Thus, in our coupled simulations, the ocean is only indirectly constrained by imposed corrections. There is, however, one important difference between our use of TBC in our coupled and uncoupled simulations: in the coupled simulations, only the fields u, v, T, and p_s are corrected with the mean increments. Specific humidity is not corrected[2].

b. The uncoupled and coupled GEOS-5 model

The results presented here are based on two different versions of the GEOS-5 model: an atmosphere-only version and a coupled atmosphere-ocean version. This allows us to address different model deficiencies and their corresponding impacts on forecast skill. More generally, it lets us assess the performance of the TBC approach within both coupled and uncoupled environments.

The uncoupled GEOS model used here is the same AGCM used to generate MERRA-2, though here the model is run at a coarser horizontal resolution (approximately 1°). As described in Gelaro et al. (2017), this AGCM includes the finite-volume dynamical core of Putman and Lin (2007), which uses a cubed sphere horizontal discretization and 72 hybrid-eta levels from the surface to 0.01 hPa. Recent upgrades to the physical parameterization schemes (in going from the original MERRA to MERRA-2) include increased re-evaporation of frozen precipitation and cloud condensate, changes to the background gravity wave drag, and an improved relationship

[1] We did not consider trying to also correct the ocean since it is unclear that the ocean analysis are of sufficient quality to estimate the necessary biases, though this may ultimately be the best approach.
[2] This was done out of an initial concern about possible negative impacts on fresh water flux into the oceans, though this has since been found to not be an issue.

between the ocean surface roughness and ocean surface stress. The model also includes a Tokioka-type trigger on deep convection as part of the Relaxed Arakawa-Schubert (RAS, Moorthi and Suarez 1992) convective parameterization scheme (Bacmeister and Stephens 2011). A new glaciated land representation and seasonally-varying sea ice albedo were implemented for MERRA-2, leading to improved air temperatures and reduced biases in the net energy flux over these surfaces (Cullather et al. 2014). The model includes the Catchment land surface model developed by Koster et al. (2000). Further details about this version of the GEOS AGCM can be found in Molod et al. (2015).

The coupled model (AOGCM) used here is part of the Subseasonal to Seasonal (S2S) prediction system that is (at the time of this writing) being used by the GMAO to provide forecasts to the North American Multi-Model Ensemble (NMME) project on a real time basis (though here our coupled model is run at coarser resolution). The model is described in more detail in Molod et al. (2018). The AGCM component of the AOGCM is a more recent version of the GEOS AGCM (described above) though it is fundamentally the same as the MERRA-2 version. The new AGCM includes parameter changes to enhance surface drag over land and oceans, to enhance form drag, and to enhance parameterized convection in the extratropics, all designed to improve weather forecast skill.

The ocean component of the GEOS AOGCM is the Modular Ocean Model version 5 (MOM5) developed at the Geophysical Fluid Dynamics Laboratory described in Griffies et al. (2005). The sea ice component is the CICE 4.1 model developed by the Los Alamos National Laboratory (Hunke and Lipscomb 2008). The ocean and atmosphere exchange fluxes of momentum, heat and fresh water through a "skin layer" interface that includes a parameterization of the diurnal cycle.

c. The experiments

The AGCM and AOGCM experiments analyzed in this study are listed in Table 1. Both models were forced with time varying GHGs as described in Appendix A of Schubert *et al.* 2014. The AGCM simulations (forced with the same observed SST and sea ice fraction used in MERRA-2[3]) consist of: (i) a long term control simulation (CNTRL-A), and (ii) a simulation (TBC-A) equivalent to CNTRL-A except for the continual correction of the model tendency biases using the TBC approach, with the correction terms (in u, v, T, q and p_s) taken directly from MERRA-2. The CNTRL-A model and the TBC-A model were also used to produce hindcasts (with observed SST) initialized from MERRA-2. In this set of hindcasts, the hindcast year's data are excluded from the estimation of the bias correction terms.

The AOGCM simulations consist of a long control simulation (CNTRL-C), a run replayed to the MERRA-2 atmosphere fields of u, v, T, and p_s (REPLAY-C), and a third run (TBC-C) in which the TBC approach is used to correct the u, v, T, and p_s tendencies, using corrections estimated from REPLAY-C. In addition, seasonal hindcasts were produced using both the CNTRL-C model and the TBC-C model, with initial conditions taken from REPLAY-C (again, with the hindcast year's data excluded from the estimation of the bias correction terms).

3. Results

[3] As summarized in Gelaro et al. (2017), the MERRA-2 SST are based on a combination of different high resolution daily NOAA OISST and OSTIA products, though prior to 01 Jan 1982, it is based on the CMIP mid-monthly 1° data.

We present here the results of applying the TBC to the GEOS model. Sections 3a and 3b focus on the impacts on the climatological biases of the AGCM and AOGCM, respectively, while Section 3c examines the impact on forecast skill.

a. TBC in the uncoupled model

The impact of TBC on the mean climate and climate variability in the AGCM is estimated from the TBC-A and CNTRL-A simulations (see Table 1). Climatological biases are defined here as long-term averaged differences from MERRA-2 and other observations as indicated below.

Figures 1 and 2 show the impact of the TBC on the zonal mean climatological biases for DJF and JJA respectively for the u-wind and specific humidity. We present, in the left panels, the climatological biases (CNTRL-A – MERRA-2), in the middle panels the improvement with TBC (TBC-A - MERRA-2), while in the right panels we show (TBC-A - CNTRL-A) to more clearly illustrate the impact of the TBC. The zonal wind biases in CNTRL-A are characterized by a poleward shift of the jets in both summer hemispheres (evident from the north/south dipole structure of the differences), with some tendency for an equator-ward shift in the winter hemispheres. In TBC-A, the poleward shift of the summer jets is substantially corrected, especially during JJA. There is less improvement in the winter jets; in fact, the SH high latitudes show, for TBC-A, an increased positive zonal wind bias during JJA (top center and right panels of Fig. 2). The reason for this is unclear but likely reflects a cold bias that develops during JJA throughout the troposphere over the SH polar regions in TBC-A. Section 4 provides a discussion of possible reasons for why the TBC does less well in correcting the climatological biases in some regions/seasons. TBC also acts to reduce substantially the zonal mean specific humidity climatological biases, especially the relatively large positive biases that occur in the lower/middle troposphere on either side of the equatorial moisture maximum during DJF (bottom panels of Fig. 1), as well as the biases in the midlevel tropics (just south of the equatorial maximum) and lower tropospheric NH middle latitudes during JJA (bottom panels of Fig. 2). We note that in both seasons the TBC-A acts to correct (strengthen) the upward motion regime of the tropics, so much so that the simulated Hadley cell is essentially indistinguishable from that in MERRA-2 (not shown).

Figure 3 shows the results for the 250mb u-wind (left column), two-meter temperature over land (T2m, middle column) and precipitation (right column) for JJA. Here again, we present in the top panels, the climatological biases (CNTRL-A – MERRA-2), in the middle panels the improvement with TBC (TBC-A - MERRA-2), while in the bottom panels we show (TBC-A - CNTRL-A) to more clearly illustrate the impact of the TBC. The impact of TBC-A is to eliminate almost completely the prevailing zonal wind climatological biases throughout the NH, especially the weak jet in the North Pacific. In the SH, where the biases are much weaker to start with, TBC-A is less effective, and in fact (as we saw in Fig. 1) generates a positive zonal wind bias at high latitudes. In the NH, the impact of TBC-A on JJA T2m is remarkable, as it eliminates most of the large positive biases, especially those over Asia and North America. The climatological precipitation biases (top right panel of Fig. 3) also show substantial improvement in many regions, with a reduction of large biases over Tibet, the maritime continent, the ITCZ, the NH storm tracks, and North America (middle right panel of Fig.3): impacts which are perhaps more clearly seen from the TBC-A – CNTRL-A fields in the bottom right panel of Fig. 3. Particularly noteworthy is the substantial reduction in the dry bias over the US Great Plains, a long-standing problem in the GEOS model and many other climate models (e.g, Lin et al. 2017).

The TBC-A impact, however, is not positive everywhere, with the increased wet bias over India being perhaps the most glaring deficiency.

We next turn our attention to the transients during JJA (Figure 4). These are based on 6 hourly data (with the monthly means removed) and include the time mean vertically-integrated zonal momentum transport ($\overline{u'v'}$, left panels), the 250mb meridional wind variability ($\overline{v'^2}$, middle panels – a measure of storm track activity [e.g., Chang and Fu 2002]), and the 850mb moisture transport ($\overline{v'q'}$, right panels). The climatological biases in all three quantities are apparent, and there are substantial corrections in the NH with TBC-A. In particular, substantial improvements are seen in the NH momentum transport, especially in the North Pacific and North Atlantic jet exit regions, where the high frequency eddies are expected to maintain the mean jet through barotropic decay (e.g., Chang et al. 2002). Also, the negative biases in $\overline{v'^2}$ (indicating weak storm tracks) seen in CNTRL-A, especially in the eastern North Pacific and the North Atlantic, are reduced in TBC-A by more than a factor of two in many places, an improvement that occurs in conjunction with the improved (strengthened) jet in these regions. Similar improvements are seen for moisture transport, with substantial increases in northward transport in the NH storm tracks in TBC-A. Also of note for TBC-A is the increased northward moisture transport over the central US, an improvement that very likely contributes to the aforementioned increased precipitation in this region. The TBC appears to be less effective in improving the JJA transients in the SH, especially over the high latitude oceans.

b. TBC in the coupled Model

We now examine the impact of applying the atmospheric TBCs obtained from REPLAY-C to the fully coupled GEOS-5 AOGCM. In assessing the impact of TBC, we compare the TBC-C results to those from both the CNTRL-C and the REPLAY-C runs. As described in Section 2, the replay approach allows us to force any model to remain close to the reanalysis during the course of an integration, providing as a byproduct the information needed to compute the TBC terms. If the model used in the replay is identical to that used to produce the original reanalysis, then one simply reproduces that reanalysis exactly. If, on the other hand, the modeling system differs from that of the reanalysis (as it does here, for three reasons: we use an updated AGCM [see above], we couple this AGCM to an ocean model, and we run at a lower resolution), identical results are not guaranteed, especially for quantities (e.g., precipitation) that are not directly constrained by the analysis increments. Given these considerations, the "replayed" results (REPLAY-C) can be considered an upper bound to what can be achieved from the TBC. Further details of the replay approach and some caveats concerning the stability of the procedure can be found in Takacs et al (2018).

Figure 5 (left panels) shows the biases for the annual mean SST. The top left panel shows that the replay approach (REPLAY-C) is able, for the most part, to reproduce the annual mean observed (Reynolds) SST. In contrast, the free-running CNTRL-C (middle left panel) shows large positive SST biases over much of the tropics and SH. These biases are essentially eliminated when TBC is applied (bottom left). In fact the performance of the TBC-C simulation is quite similar to that of REPLAY-C over much of the world's oceans. TBC-C also reduces the cold biases in the North Pacific, though not to the extent seen in REPLAY-C. While TBC-C provides little improvement in the tropical SST annual cycle (not shown), this cycle is already fairly realistic in CNTRL-C. In fact, TBC-C seems to have introduced a slightly exaggerated annual cycle in the central Pacific.

The impact of TBC-C on tropical SST variability is shown in the right three panels of Figure 5. CNTRL-C clearly has excessive variability (tied to ENSO) compared with the observations. In contrast, the variability in the TBC-C run has more reasonable amplitude, though TBC does miss very strong events of the type that occurred in nature during this time period (e.g., 1982/83, 1997/98, 2015/16)[4]. As a result, the overall SST variability in TBC-C is somewhat weaker than the observed variability.

Turning next to the results for the zonal mean atmosphere, TBC-C produces substantial reductions in the biases of the zonal mean zonal wind almost everywhere (and especially in the subtropics) for both seasons (top panels of Figs. 6 and 7). The improvement in the zonal mean specific humidity (bottom panels of Fig. 6 and 7) is also substantial, highlighted by the elimination of the wet biases in CNTRL-C in the tropics and SH during both seasons (it is noteworthy that this occurs despite not correcting the moisture). We note that the TBC-C produces little improvement in the zonal mean vertical motion during DJF (not shown) in contrast to the improvement seen in the AGCM simulations. However, there is a rather substantial improvement during JJA including a reduction in the anomalous upward motion in the upper troposphere just north of the equator.

Figure 8 shows the biases in the DJF (left panels) and JJA (right panels) precipitation for REPLAY-C (top panels), CNTRL-C (middle panels), and TBC-C (bottom panels). We see that much of the excessive precipitation that occurs just north of the equator in the Pacific during both seasons in CNTRL-C is reduced in the REPLAY-C run, as is the excessive precipitation in the tropical Atlantic and the Indian Ocean. The large dry bias over India and wet bias over

---

[4] We note that there is no reason for the simulations to have ENSO events synchronized with those in nature, though the models are run with observed CO_2 and other greenhouse gases, explaining the positive trend seen in the SST in both the observed and simulated SSTs.

Southeast Asia during JJA in CNTRL-C are also reduced in REPLAY-C. REPLAY-C does introduce a substantial dry bias over South America during DJF that is not evident in CNTRL-C. REPLAY-C also does little to reduce the dry bias over the US Great Plains; in fact it appears to exacerbate it compared to the control. Since winds and temperature in the replay are essentially the same as those in MERRA-2, the lack of improvement in the precipitation over the US Great Plains and the other regions mentioned above almost certainly reflects the fact that we do not replay the moisture in the AOGCM.

The TBC-C run produces some of the same improvements indicated above for the REPLAY-C run. The TBC-C is, however, less effective in reducing the excessive Pacific precipitation that occurs north of the equator, especially during DJF; in fact, TBC-C appears to be slightly worse than CNTRL-C in the eastern tropical Pacific, with a dry bias just south of the equator and a somewhat larger wet bias south of that. During JJA, TBC-C successfully reduces the dry bias over India, reduces the wet bias over Southeast Asia, and is somewhat more successful (compared with DJF) in reducing the excessive precipitation over the tropical Pacific.

We note that while TBC-C does produce overall more realistic atmospheric (e.g., OLR) variability in the tropics, primarily by reducing the excessive variance found in the CNTRL-C run (not shown), it does little to improve the MJO, though the CNTRL-C model already produces a fairly realistic but weaker-than-observed MJO (D. Achuthavarier, personal communication).

We next focus on DJF, with an eye towards assessing how TBC-C affects ENSO-related teleconnections over North America during that season. Since ENSO has large impacts on the North Pacific/North American jet and stationary waves, improvements in the climatologies of those aspects of the flow should have positive impacts on ENSO-related teleconnections. Figure

9 (left panels) shows that TBC-C corrects the excessive subtropical westerly winds that extend across the North Pacific, the southern United States, and the North Atlantic. It also eliminates the easterly bias in the eastern tropical Pacific. It does little to correct the relatively small biases seen for CNTRL-C in the SH. The TBC-C run also substantially improves the boreal winter stationary waves (right 3 panels of Fig. 9), particularly the position, structure and amplitude of both the ridge over the west coast of North America and the upstream trough.

Turning next to the DJF transients (Fig. 10), we see that CNTRL-C has anomalously large transient wave activity (as reflected in the 250mb kinetic energy, left panel) centered at about 30°N and generally over the continents. This bias, which is presumably linked to the excessive subtropical westerlies noted earlier, is corrected in the TBC-C run. In fact, TBC features transients that, compared to MERRA-2, are slightly too weak in the NH and, while somewhat improved, remain too weak in the SH. TBC-C shows large improvements in the NH 200mb zonal momentum flux (middle panels) and also shows improvements in the 850mb transient moisture transport (right panels), particularly just south of the storm track regions.

On interannual time scales, TBC-C primarily acts to reduce some of the excessive DJF stationary wave variance that occurs in CNTRL-C over the northeast Pacific, northern Eurasia and eastern North America (left 3 panels of Fig. 11). While these impacts are positive overall, the reduction over the eastern North Pacific results in a variability that is now somewhat too weak. The reductions are likely due to TBC-C-induced changes in the (now reduced) variability of the tropical Pacific SST linked to ENSO, which is known to contribute to the height variability over the North Pacific/North American region (e.g., Diaz et al. 2001). The interannual link between the tropical Pacific SST and the 250mb eddy height field for DJF is quantified in Figure 11 (right panels) in terms of the correlation between eddy height and the

Nino3.4 index. TBC-C shows a weakening of certain biases seen in CNTRL-C, particularly the unrealistically strong negative correlations over much of the United States and southern Canada and the strong positive correlations to the north. The overall spatial pattern of the correlations over the North Pacific/North American region is also improved.

3c. Forecast Skill

In this section we assess the degree to which TBC increases forecast skill over North America in both the uncoupled (Section 3ci) and coupled model (section 3cii). In the uncoupled case, we focus on boreal summer and subseasonal time scales, for which coupling to the ocean is likely of secondary importance. In the coupled case, we focus on boreal winter and seasonal time scales, for which ENSO is known to have an important impact on forecast skill.

i. Boreal Summer and the uncoupled model

Our focus here is on the extent to which the improvements in the subtropical/middle latitude jets and transients in the TBC-A model described in Section 3a lead to improvements in subseasonal boreal summer forecast skill over North America. The skill assessment is based on a series of hindcasts initialized in late spring and running through August produced with both the CNTRL-A and TBC-A models (see section 2c). Note that in the following we use the terminology hindcasts and forecasts interchangeably, keeping in mind that these simulations are not true forecasts; in these atmosphere-only runs, observed SSTs are prescribed throughout the forecast period.

The connection between forecast skill and the quality of a model’s climate (including variability) is not straightforward, though it seems plausible that a model with a better long-term climate should have better forecast skill. Even if that is the case, correcting climate drift (which is a function of forecast lead time, see Section 2a) can presumably only lead to improved forecast skill if a substantial amount of the bias (and its correction) occurs before all predictability is lost. Therefore, these two time scales (associated with drift development and predictability) serve to define a window of forecast leads during which TBC can be expected to have an impact on skill. For example, if it turns out that it takes 3 months for the drift in the CNTRL to fully develop into the long-term climate bias, and if the underlying predictability limit is 20 days, it is unlikely that any small correction (made by TBC) to the still small bias in the CNTRL during the first 20 days would have much impact on forecast skill. In order to help address this issue we decompose the total mean square error (MSE) into the following terms:

$$\overline{\langle (F-O)^2 \rangle} \;=\; \overline{\left[\left(\langle F\rangle - \overline{\langle F\rangle}\right) - (O-\bar{O})\right]^2} \;+\; \overline{\langle (F-\langle F\rangle)^2 \rangle} \;+\; \left(\overline{\langle F\rangle} - \bar{O}\right)^2, \qquad (1)$$

where the angle brackets denote an ensemble mean and the over-bar a time mean; also F denotes a forecast and O denotes the observations (MERRA-2). The first term on the right hand side (RHS) is the MSE after first removing the respective time means. We will refer to this term as the unbiased MSE. The second term on the RHS is the MSE of a perfect model (the ensemble mean predicting one ensemble member), and the third term is the MSE associated with the climate drift. This latter term quantifies the evolution of the bias or drift as discussed earlier, saturating at long leads (when the forecast has lost all memory of the initial conditions) to the square of the climatological bias.

The top left panel in Fig. 12 shows the decomposition for the 250mb u-wind (in terms of RMSE, averaged over the NH) [5], and the middle left panel shows the decomposition for the 250mb v-wind (averaged over the middle latitude North Pacific). These two quantities should give us a sense for how the drift in the wave-guide evolves (u250mb) and the extent to which the Rossby waves themselves are predicted more accurately (v250mb). Both the u-wind and the v-wind total errors (blue curves) saturate by about 15 days (slightly longer for the u-wind) regardless of whether the model is corrected or not. This reflects the underlying predictability limits of the model (red curves), which is about 20 days. The unbiased RMSE (black curves) indicate no improvement in the v-wind TBC-A skill compared to CNTRL-A by this metric. The bottom left panel of Fig. 12 shows that there is, however, apparently some very modest improvement in the correlation beginning somewhat before day 10, though this occurs only after the skill for both CNTRL-A and TBC-A is rather small (about 0.3). To assess whether these averaged results represent significant improvements we show in the right panels of Fig. 12 an example of the spatial distribution of the correlations at a lead of 12 days. The differences show generally positive values with statistically significant improvements along the storm track – the region we would expect to see improvements in light of the improved North Pacific jet. We note that in comparison, the perfect model correlations are substantially larger than the actual correlations, (above 0.4 at 10 days), suggesting that further improvements in skill may be possible.

The fact that the improvement in the v-wind is modest and doesn't occur until after the first week in the forecasts likely reflects the fact that the bias in the u-wind (the wave guide) develops

[5] The values are obtained by first computing the MSE at each gridpoint. These values are then averaged over the indicated regions, after which the square root is taken to obtain the RMSE. Correlations are computed similarly with the covariances computed at each grid point and then averaged over the indicated regions.

slowly over the course of about 2 months (green curves in top left panel of Figure 12). As a result, any impacts on T2m and precipitation forecast skill over North America from the improvements in the wind forecasts are likely confined to week two (after that the v-wind skill is likely too low (< 0.1) to have an impact. Ultimately the longest lead times at which we can expect some improvement are constrained by the v-wind limit of predictability, which is about 20 days. Having said that, we find essentially no improvement in North American precipitation forecasts with TBC at those leads (not shown). On the other hand, we do find some improvement in T2m forecasts (Fig. 13), especially when we condition the forecasts on the amplitude of the leading Rossby wave impacting North American climate in summer (lower left panel of Fig 13; see also Schubert et al, 2011). The results shown for day 10 (top right panel of Fig. 13) indicate that some of the largest improvements occur over Canada, consistent with where we expect the leading Rossby wave to have the greatest impact on T2m (lower right panel of Fig 13). This increased skill in predicting T2m apparently reflects the fact that the leading RCEOF is forecast with greater skill in the TBC-A hindcasts after the first week (not shown).

ii. Boreal Winter and the Coupled Model

Our focus here is on whether the TBC approach applied to the coupled model leads to improved boreal winter seasonal forecasts, especially over North America, where we expect that any improvements in SST variability, stationary waves, and ENSO-related teleconnections might translate into improved forecast skill. The forecasts were initialized on 1 November of 1985-2015 and consist of 10 ensemble members for both the CNTRL-C and TBC-C models (see Section 2c).

Figure 14 (top two panels) again provides an integrated overview of the coupled hindcast results decomposed into the various terms of (1), focusing in this case on the eddy 250mb height field. Averaged over the NH (top panel), the total error appears to saturate after about two months (in January), with CNTRL-C showing larger total error than TBC-C. The larger total variance of the CNTRL-C appears to reflect an intrinsic property of the models (evident in the perfect model results – the red curves), but it is also in part due to the development of a larger bias in the control (green curves) both early in the forecast (November/early December) and again starting in January. The perfect model RMSE approaches the unbiased RMSE (black curves) by mid-December, indicating that most of the predictability (based on RMSE) is lost by that time. In early February there is some hint that the TBC-C predictions have somewhat smaller unbiased RMSE than the control (black curves). For comparison, the results for the SH indicate little difference between the TBC-C and CNTRL-C hindcasts in either the drift or RMSE, consistent with the less substantial TBC-derived improvement to the SH climate.

The bottom panel of Figure 14 shows the correlations with MERRA-2 for the PNA region (150°E-300°E, 20°N-80°N). The perfect model and TBC-C correlations with MERRA-2 both drop to 0.3 by the beginning of December, the time at which the NH RMSE approaches saturation. The control correlations with MERRA-2 drop even faster, reaching 0.2 by this time. There is however, some indication of a return of skill during January and February in the perfect model results, presumably linked to the stronger impact of ENSO over some parts of North America during these months (e.g., Jong et al. 2016). The return of skill is also evident in the TBC-C hindcasts, though less so in the control hindcasts. The apparent increase in skill during February is consistent with Chen et al. (2017), who found that, for ENSO-related T2m and precipitation predictions over North America, the skill for all of the NMME models tended to be higher in February than in other winter months.

The evolution of the climate drift in the 250mb eddy height fields (and the correlations) shown in Figure 14 suggests that any improvement in wintertime seasonal forecast skill from TBC-C over North America is likely to occur early on (during the first month of the forecast) and late in the forecast at lead times beyond roughly 2 months.

The left set of 9 plots in Figure 15 show the hindcast skill of the 250mb eddy height over the Pacific/North American region averaged over the early (16Nov-15Dec), middle (16Dec-15Jan), and later (21Jan-01Mar) segments of the predictions. The correlations (with MERRA-2) are shown for the CNTRL-C (middle row) and TBC-C (top row) hindcasts; differences are shown in the bottom row. The correlations in both sets of hindcasts are overall, as expected, high over the tropics/subtropics, with some relatively high correlations (>0.6) also occurring over the North Pacific, western North America and the southeast US. Over North America, the difference maps show some improvement in skill for the TBC-C height hindcasts for the early segment and again some improvement for the late segment (though marginally significant), with no improvement for the middle segment – results that are consistent with the line plots of the correlations in Fig. 14. These apparent improvements in the skill of the eddy height predictions occur in the absence of any significant improvements in the tropical Pacific SST forecasts (not shown).

The middle (right) set of 9 panels of Figure 15 show the correlations for T2m (precipitation) over North America. As with the eddy heights, the largest improvements for T2m hindcasts occur early on and again late in the forecasts, with no skill, or even reduced skill, compared to CNTRL-C for the interval in between. For precipitation, TBC shows overall little improvement in skill, with some scattered improvements along the west coast early in the forecast. During the middle period, TBC-C actually shows substantial areas of degraded skill relative to CNTRL-C especially over the southeastern US.

4. Discussion and Conclusions

This study examined the overall impact of correcting biases in short-term atmospheric tendencies in a general circulation model. Results are presented for two different versions of the NASA/GMAO GEOS model (an AGCM forced with observed SST, and an updated AGCM coupled to an ocean model). Our experiments show that state-independent TBC to the atmosphere can produce considerable improvements to the simulated mean climate as well as to its variability on subseasonal and, to some extent, seasonal and longer time scales. The improvements are, however, not uniform and depend to some degree on the quantity, region, and season, as well as the model itself.

In discussing the TBC impacts on the model's climate, it is useful to consider them as being divided into those that are direct and those that are indirect, with the latter including any quantities (such as precipitation and, for the AOGCM, atmospheric moisture) that are not explicitly forced by the TBC, as well as the transients, since the TBC is a constant forcing term. It should be emphasized, however, that even for those quantities directly forced by the TBC (e.g., u, v, T), it is not a forgone conclusion that the tendency errors in these terms will be fully corrected by constant forcing terms. There are several possible reasons for this including the possibility that the true errors cannot be represented by a simple constant forcing term and are in fact state dependent (e.g., Leith 1978, Danforth et al. 2007), as well as the possibility that, even if the errors can be represented in that way, the TBCs may be poor estimates of the true corrections as a result of statistical sampling errors and/or as a result of deficiencies/biases in the reanalysis. Furthermore, it is not obvious that a model will respond to the increments in a physically realistic way. It is quite possible that, for example, correcting the moisture and temperature profiles would lead to spurious feedbacks from the model's convective scheme, which may have been tuned to produce realistic precipitation with somewhat different profiles. In the following, we provide some examples from our results that serve to illustrate these issues.

The improvements in the middle latitude transients in both the AGCM and AOGCM are a clear example of a positive indirect impact – an impact that is very likely strongly tied to the improvements in the jets. The nature of the improvements in the jets (or lack of improvement in some cases) appears to vary with the seasons, the hemisphere, and the model in question. TBC-A corrects the poleward shift of both summer jets, consistent with increased drag on the jets (e.g., Robinson 1997). Since the summer jets are largely eddy driven (e.g., Lachmy and Harnik, 2016), it is likely that the improvement in the jets also drives (and interacts with) the improved transient eddy momentum transport. Additional work (not shown) indicates that jet biases throughout the Northern Hemisphere are particularly sensitive to temperature errors over and near Tibet, suggesting that corrections in this area may be especially important in correcting the NH summer jets (an example of a positive indirect impact). The TBC-A does less well in correcting the high latitude zonal winds in the SH upper troposphere/lower stratosphere during winter, suggesting that uncorrected errors in stratospheric dynamics and reanalysis quality (poor estimates of the increments) may be issues.

The primary zonal wind errors in the AOGCM appear to be fundamentally different in character compared with the AGCM errors, consisting of excessive subtropical westerlies in both hemispheres (though more so in the NH) and during both seasons. These likely reflect anomalous forcing/heating by the excessively strong and split ITCZ in the coupled model. The fact that the TBC-C corrects these zonal wind errors (and associated transients), yet makes only modest corrections to the tropical precipitation (especially during DJF), indicates that the corrections to the zonal wind errors are forced more directly by the increments. In fact, it appears that it is the tropical mid-tropospheric temperature increments that appear to play a key role during DJF, presumably in part by reducing the strong tropical warm bias in that run. At longer time scales, the impacts on the variability of the SST (and the associated changes in tropospheric height variability) in the TBC-C run is likely tied to improvements in the equatorial surface stress (not shown), though exactly how that acts to reduce the ENSO variability is unclear. We note that the dramatic reduction of the SST bias in TBC-C appears to be the result of a combination of direct impacts from the near surface temperature increments (especially over the Gulf Stream, the SH high latitudes, and equatorial and coastal upwelling regions) and indirect impacts due to the reductions in surface stress biases.[6]

Perhaps the strongest test of the TBC for improving the climate characteristics of the model is the extent to which the components of the hydrological cycle are improved. We have seen clear improvements in the precipitation in the AGCM results, both in the tropics and in the US Great Plains. Also, improved (increased) cloudiness in TBC-A (not shown) appears to contribute to the dramatic reduction in the warm bias over the NH summer continents. Here we have a clear case where the TBC impacts are indirect; the model's parameterizations of moisture processes working with the states directly affected by TBC appear to produce more realistic output – a result likely helped by the fact that the AGCM is the same as that used to generate

[6] The bias in cloud fraction has actually increased in TBC-C (less cloudiness), indicating this did not contribute to the reduction in the SST warm bias.

MERRA-2 (though run at lower resolution). In contrast, TBC-C produced considerably less improvement to the precipitation, including little improvement to the ITCZ (especially during DJF), but also no improvement to the summer dry bias over the US Great Plains. Here it is instructive to compare the TBC-C and REPLAY-C runs. To a large extent the lack of improvement (or even degradation as seen over South America in DJF) in the TBC-C run is already reflected in REPLAY-C run. As such, this does not appear to reflect a limitation of the TBC approach, but instead an inconsistent or lack of forcing by the increments (recall that we don't correct the moisture in the AOGCM).

A key goal of this study was to determine whether the improved climate characteristics of the model induced by TBC translate into improved forecast skill (perhaps the ultimate indirect impact). We found, however, that TBC-related skill improvements were rather modest at best at both subseasonal and seasonal time scales. For the uncoupled case, where our focus was on boreal summer and subseasonal forecasts, the improvements in the NPSJ and the transient eddy activity led to only modest improvements in the T2m forecasts over North America (and only when conditioned on the leading Rossby wave impacting North America), and to no improvement in the precipitation forecasts. In the coupled case, our focus was on improving boreal winter forecast skill over North America at seasonal time scales. Here too, despite various improvements to the stationary waves and related transients, and despite what appear to be more realistic ENSO variability and associated teleconnections, the impact of TBC on skill was not uniform with forecast lead and was again overall quite modest.

We interpret these hindcast results in terms of predictability limits and the time it takes the relevant aspects of climate drift to become large enough to begin having an impact on skill (and thus the time it would take for TBC-based reductions of the drift to affect the skill). In the uncoupled case, focusing on boreal summer and North America, the climate drift in the North Pacific waveguide (believed to be a key controlling factor for Rossby waves entering North America) appears to develop too slowly in CNTRL-A (reaching only about ½ the long term value at 10 days lead) to allow its correction in TBC-A to produce more than a modest impact (via more skillful Rossby wave predictions) on week-two T2m forecasts (when skill is already rather low). In the coupled case, focusing on boreal winter over North America, our assessment of the drift in the stationary waves suggests two adjustment time scales: an early drift that develops during the first month (presumably dynamically driven) and a more slowly developing drift that occurs during months 3 and 4 (presumably linked to deficiencies in coupled processes). In contrast, the corrected model experienced an early drift that took longer to develop than in the control, and never experienced the slow drift of the control model during months 3 and 4. There thus appear to be two windows (one early and one late) during which TBC could induce improved forecasts. This indeed appears to be borne out in the forecasts of both eddy heights over the Pacific/North American region and T2m over North America.

Additional improvements in forecast skill might be possible with a state-dependent correction if the associated statistical sampling issues can be overcome (e.g., Leith 1978; Danforth et al. 2007). In fact, it is possible that the modest impacts on skill (or even reductions in skill) found here reflect the presence of state-dependent errors that may or may not be in phase with the state-independent errors. Our TBC approach nevertheless provides a reasonable baseline of what can currently be achieved with state-independent corrections to a global climate model employing a state-of-the-art atmospheric reanalysis. The approach is relatively easy-to-implement and, since it is based on very short-term forecasts when the error growth is still linear, appears to produce corrections that (to a large extent) reflect physically realistic adjustments to the model equations.

It is however likely that substantial further improvements will require model system improvements not directly addressed by TBC, improvements involving, for example, land/atmosphere interaction, cloud/radiative processes, and initialization procedures for (and quality of) atmospheric, land and ocean states. While potential improvements in forecast skill may not be the main impetus for carrying out the TBC, we believe that TBC-induced improvements in transients, stationary waves, and other climate characteristics can be a key motivating factor for employing the approach. Such improvements can make the model better suited for addressing a host of climate problems, such as those that require addressing regional impacts of global climate variability and change.

Acknowledgements

This work was supported by the NASA MAP (NNG17HP01C and WBS 802678.02.17.01.33), and NOAA MAPP (NA14OAR4310221) programs.

Appendix A: Methodology

As described in Takacs et al (2018), the replay approach takes advantage of the incremental analysis update (IAU) procedure employed in the GEOS data assimilation system to force a model to track a pre-existing analysis. The basic approach is shown schematically in Figure A1. The blue arrows indicate that the replay is essentially a continuous model simulation that is driven by a sequence of IAU forcing terms (updated every 6 hours) computed as the difference between a short forecast and the corresponding analysis. The general form of the equations governing a replay can be written (for a quantity q) as:

$$\frac{\partial q}{\partial t} = f(q) + \Delta q \qquad \text{(A1)},$$

where $\Delta q = (analysis - forecast)/6hrs$, and f(q) is the tendency associated with all the dynamics and physics terms of the model – in other words, it corresponds to the uncorrected model. For the coupled model replay performed as part of this study, the increments are computed for the winds, temperature and surface pressure.

The governing equations for the TBC approach have the same form as (1), except that the forcing term associated with the increments is no longer an instantaneous value (specific to a particular 6 hour period), but is instead a long term mean. In particular,

$$\frac{\partial q}{\partial t} = f(q) + \overline{\Delta q} \qquad \text{(A2)},$$

where the $\overline{\Delta q}$ are 6-hourly values that are averaged over the years 1980-2015 separately for each 6-hr time period of each day-of-year[7], and as such retain the diurnal and annual cycles. The

[7] In the case of the coupled model we further apply a 7-day running mean to the increments, though this is done in a way that retains the mean diurnal cycle.

above indicates that the model with the TBC (A2) can be considered as an approximation of (A1), in which the correction term is simplified to retain only the first moment statistics (the mean) of the increments: the assumption being that such simple corrections nevertheless represent physically realistic systematic adjustments to the model's physics and/or dynamics tendency terms.

As noted in the text, in the case of the AGCM, instead of replaying to MERRA-2 to obtain the Δq terms, we take advantage of the fact that the AGCM used here is the same as that used to generate MERRA-2 (though run at lower resolution) and so we take the increments directly from the MERRA-2 archive (appropriately averaged to the reduced resolution of the AGCM). This was not the case for the AOGCM. We could not simply couple the corrected AGCM to the ocean, but found it necessary to replay to the MERRA-2 atmosphere running in coupled mode to obtain the increments appropriate for correcting the biases that develop in the coupled model.

Finally, in assessing the quality of the climates of the TBC simulations, the above makes it clear that the most fair comparison to make is with the climate of the corresponding replay run, as we do for the coupled model. In the case of the AGCM (which is a lower resolution version of the same model used to produce MERRA-2) such a comparison is, however, essentially equivalent to comparing with MERRA-2, since (A1) would to a large extent reproduce the reanalysis, though of course at lower resolution.

References:

Achuthavarier D., H. Wang, S. Schubert and M. Sienkiewicz, 2017: Impact of DYNAMO observations on the NASA GEOS-5 reanalysis and the representation of the MJO initiation over the tropical Indian Ocean. J. Geophys. Res.-Atmos. 122, 179–201, doi:10.1002/2016JD025363.

Bacmeister, J. T., and G. L. Stephens, 2011: Spatial statistics of likely convective clouds in CloudSat data, J. Geophys. Res., 116, D04104, doi:10.1029/2010JD014444.

Bhargava, K., E. Kalnay, J.A. Carton and F. Yang, 2018: Estimation of systematic errors in the GFS using analysis increments. J. Geophys. Res.: Atmospheres, 123, 1626 – 1637. https://doi.org/10.1002/2017JD027423.

Bloom, S.C., L.L. Takacs, A.M. da Silva, and D. Ledvina, 1996: Data Assimilation Using Incremental Analysis Updates. *Mon. Wea. Rev.*, 124, 1256–1271, https://doi.org/10.1175/1520-0493(1996)124<1256:DAUIAU>2.0.CO;2.

Cannon, A.J., 2016: Multivariate Bias Correction of Climate Model Output: Matching Marginal Distributions and Intervariable Dependence Structure. J. Climate, 29, 7045–7064, https://doi.org/10.1175/JCLI-D-15-0679.1

Chang, E. K. M. and Y. Fu, 2002: Interdecadal variations in Northern Hemisphere winter storm track intensity. J. Climate,15, 642–658.

Chang, E.K., S. Lee, and K.L. Swanson, 2002: Storm Track Dynamics. *J. Climate*, 15, 2163–2183, https://doi.org/10.1175/1520-0442(2002)015<02163:STD>2.0.CO;2

Chang, Y., S. D. Schubert, S. J. Lin, S. Nebuda, B.-W. Shen, 2001. The Climate of the FVCCM-3 Model. NASA *Technical Report Series on Global and Data Assimilation 104606*, Vol. 20. M. Suarez, editor, Document (8273 kB).

Chen, L., H. van den Dool, E. Becker, and Q. Zhang, 2017: ENSO Precipitation and Temperature Forecasts in the North American Multimodel Ensemble: Composite Analysis and Validation. *J. Climate*, 30, 1103–1125, https://doi.org/10.1175/JCLI-D-15-0903.1

Cullather, R. I., S. M. J. Nowicki, B. Zhao, and M. J. Suárez, 2014: Evaluation of the surface representation of the Greenland Ice Sheet in a general circulation model. *J. Climate*, 27, 4835–4856, doi:https://doi.org/10.1175/JCLI-D-13-00635.1.

Danforth, C.M., E. Kalnay, and T. Miyoshi, 2007: Estimating and Correcting Global Weather Model Error. *Mon. Wea. Rev.*, 135, 281–299, https://doi.org/10.1175/MWR3289.1.

DelSole, T., and A. Y. Hou, 1999: Empirical correction of a dynamical model. Part I: Fundamental issues. Mon. Wea. Rev., 127, 2533–2545.

Diaz, H. F., M. P. Hoerling, and J. K. Eischeid, 2001: ENSO variability, teleconnections and climate change. *Int. J. Climatol.*, 21, 1845–1862, doi:https://doi.org/10.1002/joc.631

Flato, G., J. Marotzke, B. Abiodun, P. Braconnot, S.C. Chou, W. Collins, P. Cox, F. Driouech, S. Emori, V. Eyring, C. Forest, P. Gleckler, E. Guilyardi, C. Jakob, V. Kattsov, C. Reason and M. Rummukainen, 2013: Evaluation of Climate Models. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change[Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M.

Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Gelaro, R., W. McCarty, M.J. Suárez, R. Todling, A. Molod, L. Takacs, C.A. Randles, A. Darmenov, M.G. Bosilovich, R. Reichle, K. Wargan, L. Coy, R. Cullather, C. Draper, S. Akella, V. Buchard, A. Conaty, A.M. da Silva, W. Gu, G. Kim, R. Koster, R. Lucchesi, D. Merkova, J.E. Nielsen, G. Partyka, S. Pawson, W. Putman, M. Rienecker, S.D. Schubert, M. Sienkiewicz, and B. Zhao, 2017: The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *J. Climate*, 30, 5419–5454, https://doi.org/10.1175/JCLI-D-16-0758.1.

Griffies, S. M., Gnanadesikan, A., Dixon, K. W., Dunne, J. P., Gerdes, R., Harrison, M. J., Rosati, A., Russell, J. L., Samuels, B. L., Spelman, M. J., Winton, M., and Zhang, R.: Formulation of an ocean model for global climate simulations, Ocean Sci., 1, 45-79, https://doi.org/10.5194/os-1-45-2005, 2005.

Huang, B., V.F. Banzon, E. Freeman, J. Lawrimore, W. Liu, T.C. Peterson, T.M. Smith, P.W. Thorne, S.D. Woodruff, and H. Zhang, 2015: Extended Reconstructed Sea Surface Temperature Version 4 (ERSST.v4). Part I: Upgrades and Intercomparisons. *J. Climate*, **28**, 911–930, https://doi.org/10.1175/JCLI-D-14-00006.1

Hunke, C. E. and W. Lipscomb, 2010: CICE: The Los Alamos sea ice model documentation and software user's manual version 4.0 LA-CC-06-012. Tech. Rep. LA-CC-06-012

IPCC, 2013: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y.

Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1535 pp, doi:10.1017/CBO9781107415324.

Jong, B., M. Ting, R. Seager, N. Henderson, and D.E. Lee, 2018: Role of Equatorial Pacific SST Forecast Error in the Late Winter California Precipitation Forecast for the 2015/16 El Niño. *J. Climate*, 31, 839–852, https://doi.org/10.1175/JCLI-D-17-0145.1

Kirtman, Ben, Dughong Min, Johnna M. Infanti; James L. Kinter III; Daniel A. Paolino; Qin Zhang; Huug van den Dool; Suranjana Saha; Malaquias Pena Mendez; Emily Becker; Peitao Peng; Patrick Tripp; Jin Huang; David G. DeWitt; Michael K. Tippett; Anthony G. Barnston; Shuhua Li; Anthony Rosati; Siegfried D. Schubert; Michele Rienecker; Max Suarez; Zhao E. Li; Jelena Marshak; Young-Kwon Lim; Joseph Tribbia; Kathleen Pegion; William J. Merryfield; Bertand Denis; Eric F. Wood, 2014: The North American Multi-Model Ensemble (NMME): Phase-1 Seasonal to Interannual Prediction, Phase-2 Toward Developing Intra-Seasonal Prediction. *Bull. Am. Meteorol. Soc. 95 585–601*.

Koster, R. D., M. J. Suarez, A. Ducharne, M. Stieglitz, and P. Kumar, 2000: A catchment-based approach to modeling land surface processes in a general model: 1. Model structure. *J. Geophys. Res*., 105(D20), 24809-24822, doi:10.1029/2000JD900327.

Lachmy, O. and N. Harnik, 2016: Wave and Jet Maintenance in Different Flow Regimes. *J. Atmos. Sci.*, 73, 2465–2484, https://doi.org/10.1175/JAS-D-15-0321.1

Leith, C. E., 1978: Objective methods for weather prediction. Annu. Rev. Fluid Mech., 10, 107–128.

Lin, Y., W. Dong, M. Zhang, Y. Xie, W. Xue, J. Huang, Y. Luo, 2017: Causes of model dry and warm bias over central U.S. and impact on climate projections. Nature Communications, 1-8, 10.1038/s41467-017-01040-2.

Molod, A. M., L. Takacs, M. Suarez, and J. Bacmeister, 2015: Development of the GEOS-5 atmospheric general circulation model: evolution from MERRA to MERRA2. *Geosci. Model Dev*., 8, 1339-1356, doi:10.5194/gmd-8-1339-2015.

Molod, A., E. Hackert, L. C. Andrews, R. M. Kovach, B. Zhao, Z. Li, Y.-K. Lim, A. Borovikov, G. Vernieres, J. Marshak, R. Koster, J. Carton, R. Cullather, D. Barahona, S. Schubert, D. Achuthavarier, Y. Vikhliaev, J. L. M. Freire, K. M. Longo, 2018: GEOS-S2S Version 2: The GMAO high resolution coupled model and assimilation system for seasonal prediction. To be submitted to J. Geophysical Research.

Moorthi, S., and M. J. Suarez, 1992: Relaxed Arakawa-Schubert: A parameterization of moist convection for general circulation models. *Mon. Wea. Rev*., 120, 978-1002.

Orbe, C., L.D. Oman, S.E. Strahan, D.W. Waugh, S. Pawson, L.L. Takacs, and A.M. Molod, 2017: Large-scale atmospheric transport in GEOS replay simulations. *J. Adv. Model. Earth Syst.*, **9**, no. 7, 2545-2560, doi:10.1002/2017MS001053.

Putman, W., and S.-J. Lin, 2007: Finite-volume transport on various cubed-sphere grids. *J. Comput. Phys.*, 227, 55–78, doi:https://doi.org/10.1016/j.jcp.2007.07.022.

Robinson, W. A., 1997: Dissipation dependence of the jet latitude. J. Climate, 10, 176–182.

Saha, S., 1992: Response of the NMC MRF model to systematic error correction within integration. Mon. Wea. Rev., 120, 345–360.

Schemm, J.-K., and A. Faller, 1986: Statistical corrections to numerical predictions. Part IV. Mon. Wea. Rev., 114, 2402–2417.

Schubert, S., H. Wang, and M. Suarez, 2011: Warm Season Subseasonal Variability and Climate Extremes in the Northern Hemisphere: The Role of Stationary Rossby Waves, J. Climate, 24, 4773-4792.

Schubert, S., H. Wang, R. Koster, M. Suarez and P. Groisman 2014: Northern Eurasian Heat Waves and Droughts. *Journal of Climate*. 27, 3169–3207.

Seager, Richard , Martin Hoerling, Siegfried Schubert, Hailan Wang, Bradfield Lyon, Arun Kumar, Jennifer Nakamura, and Naomi Henderson, 2015: Causes of the 2011–14 California drought, *J. Climate*, 28, 6997–7024. doi: http://dx.doi.org/10.1175/JCLI-D-14-00860.1, 2015.

Seager, R. and N. Henderson, 2016: On the Role of Tropical Ocean Forcing of the Persistent North American West Coast Ridge of Winter 2013/14. J. Climate, 29, 8027–8049, https://doi.org/10.1175/JCLI-D-16-0145.1

Takacs, L.L., M.J. Suárez, and R. Todling, 2018: The Stability of Incremental Analysis Update. *Mon. Wea. Rev.*, **146**, 3259–3275, https://doi.org/10.1175/MWR-D-18-0117.1

Wang, H. and S. Schubert, 2014: Causes of the Extreme Dry Conditions over California during Early 2013, *Bull. Amer. Meteor. Soc.*, vol. 95, no. 9, Sept. 2014, p. S7-S11, 2014.

Wang, H., S.D. Schubert, and R.D. koster, 2017: North American Drought and Links to Northern Eurasia: The Role of Stationary Rossby Waves. Climate Extremes:Mechanisms and Potential Prediction, S. Wang et al. (Eds.), AGU Monograph, American Geophysical Union.

Xue, H.-L., Shen, X.-S., & Chou, J.-F. (2013). A forecast error correction method in numerical weather prediction by using the recent multiple-time evolution data. Advances in Atmospheric Sciences, 30(5), 1249–1259. https://doi.org/10.1007/s00376-013-2274-1.

Yu H., J. Huang, W. Li, and G. Feng, 2014a: Development of the analogue-dynamical method for error correction of numerical forecasts. J. Meteor. Res., 28(5), 934-947, doi: 10.1007/s13351-014-4077-4.

Yu H., J. Huang, and J. Chou, 2014b: Improvement of medium-range forecasts using the analogue-dynamical method. Mon. Wea. Rev., 142(4), 1570-1587, doi: http://dx.doi.org/10.1175/MWR-D-13-00250.1

Table 1. A summary of the AGCM and AOGCM experiments.

Exp. #	Exp. Name	Description	Model	SST
1	CNTRL-A simulation	36- year control simulation for the period 1980-2015	AGCM without TBC	observed
2	TBC-A simulation	36- year TBC simulation for the period 1980-2015	AGCM with TBC	observed
3	CNTRL-A hindcasts	hindcasts initiated every day from May 1-June 30 and run through the end of August for 1988, 1998 and 2000-2015	AGCM without TBC	observed
4	TBC-A hindcasts	hindcasts initiated every day from May 1-June 30 and run through the end of August for 1988, 1998 and 2000-2015	AGCM with TBC	observed
5	CNTRL-C simulation	36- year control simulation for the period 1981-2016	AOGCM without TBC	predicted
6	REPLAY-C simulation	36- year replay to MERRA-2 for the period 1981-2016	AOGCM replayed to MERRA-2	predicted
7	TBC-C simulation	36- year simulation with TBC for the period 1981-2016	AOGCM with TBC	predicted
8	CNTRL-C hindcasts	10-member ensemble hindcasts initialized every November 1 and run through April 1 of the following year for 1985-2015	AOGCM without TBC	predicted
9	TBC-C hindcasts	10-member ensemble hindcasts initialized every November 1 and run through April 1 of the following year for 1985-2015	AOGCM with TBC	predicted

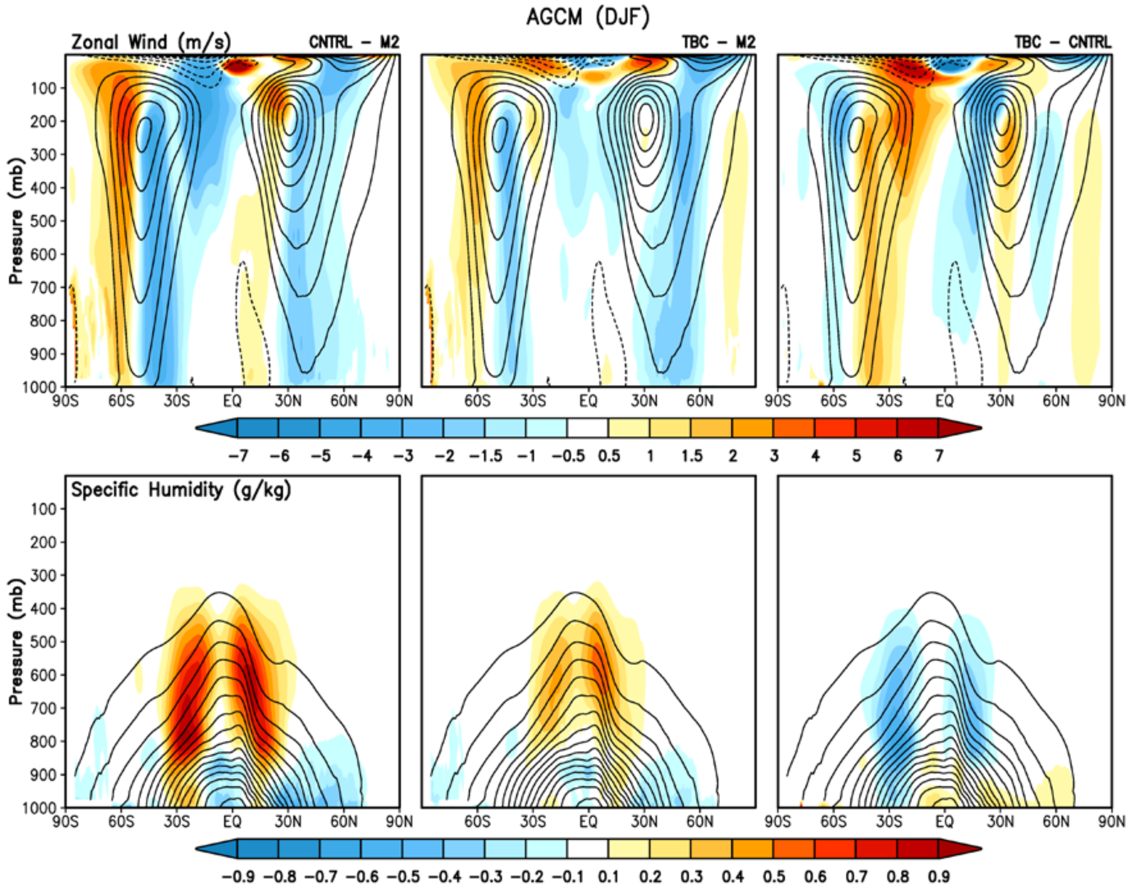

Figure 1: The zonal mean u-wind (top panels, m/s) and specific humidity (bottom panels, g/kg). Left panels: the shading indicates CNTRL-A – MERRA-2 with the climatological MERRA-2 wind fields contoured every 5 m/s in the top panels, and the MERRA-2 climatological specific humidity contoured every 1 g/kg in the bottom panels. Middle panels are the same as the left panels, except for TBC-A - MERRA-2. Right panels are the same as the left two panels, except for TBC-A – CNTRL-A. All fields are averaged for DJF over the years 1980-2015.

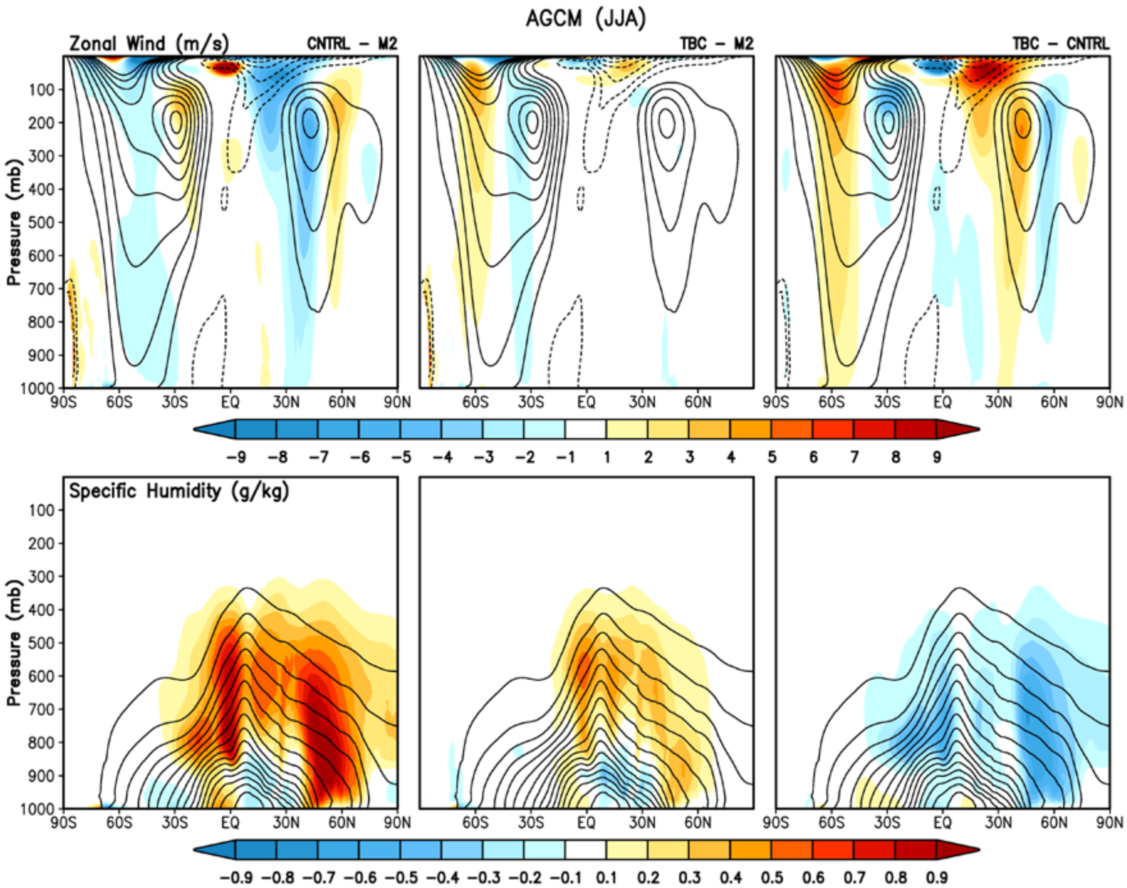

Figure 2: Same as Fig. 1, except for JJA.

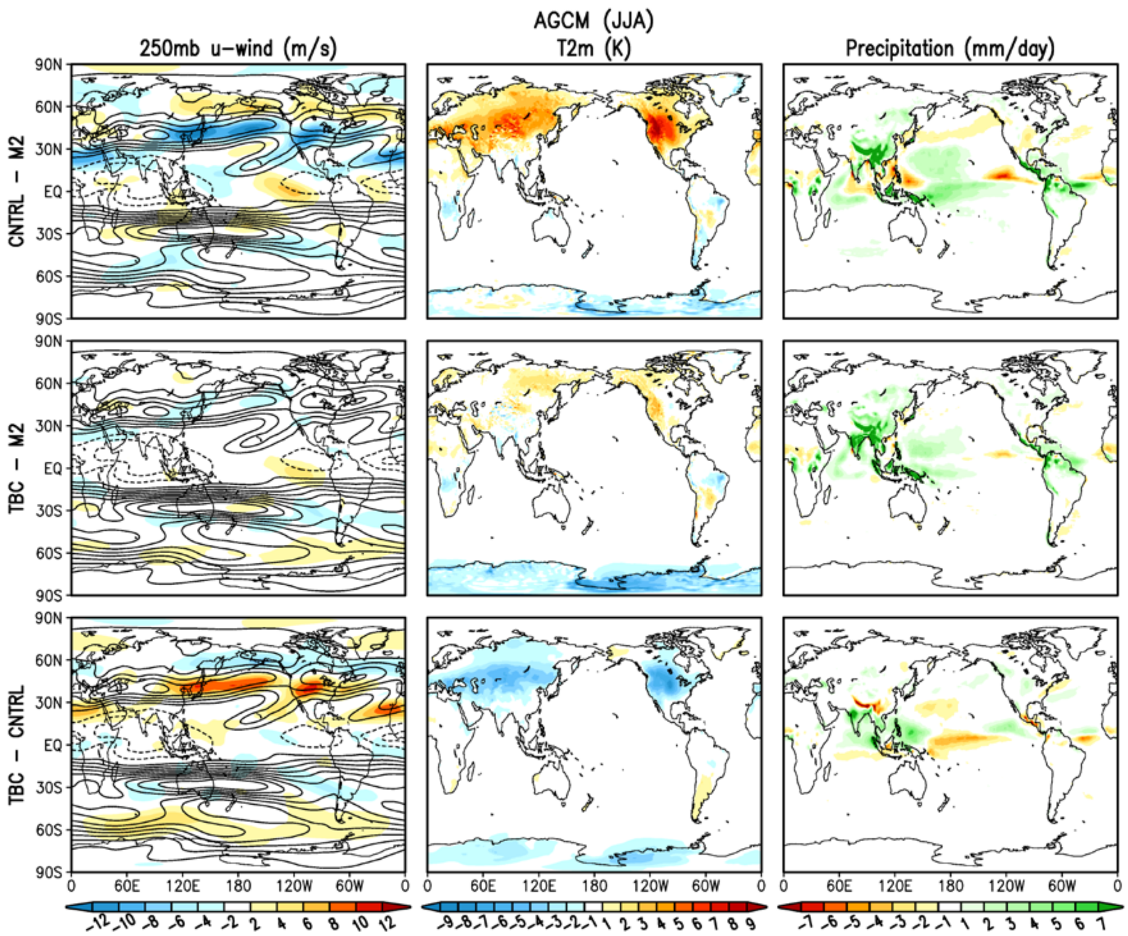

Figure 3: The 250mb zonal wind (m/s, left column of panels), two-meter temperature (°K, middle column of panels), and precipitation (mm/day, right column of panels). The shading indicates the CNTRL-A – MERRA-2 in the top row of panels, TBC-A - MERRA-2 in the middle row of panels, and TBC-A – CNTRL-A in the bottom row of panels. In the left panels the contours indicate climatological mean 250mb zonal winds from MERRA-2 (contoured every 5 m/s). All fields are averaged for JJA over the years 1980-2015. The MERRA-2 precipitation is an observationally-corrected field (Gelaro et al. 2017).

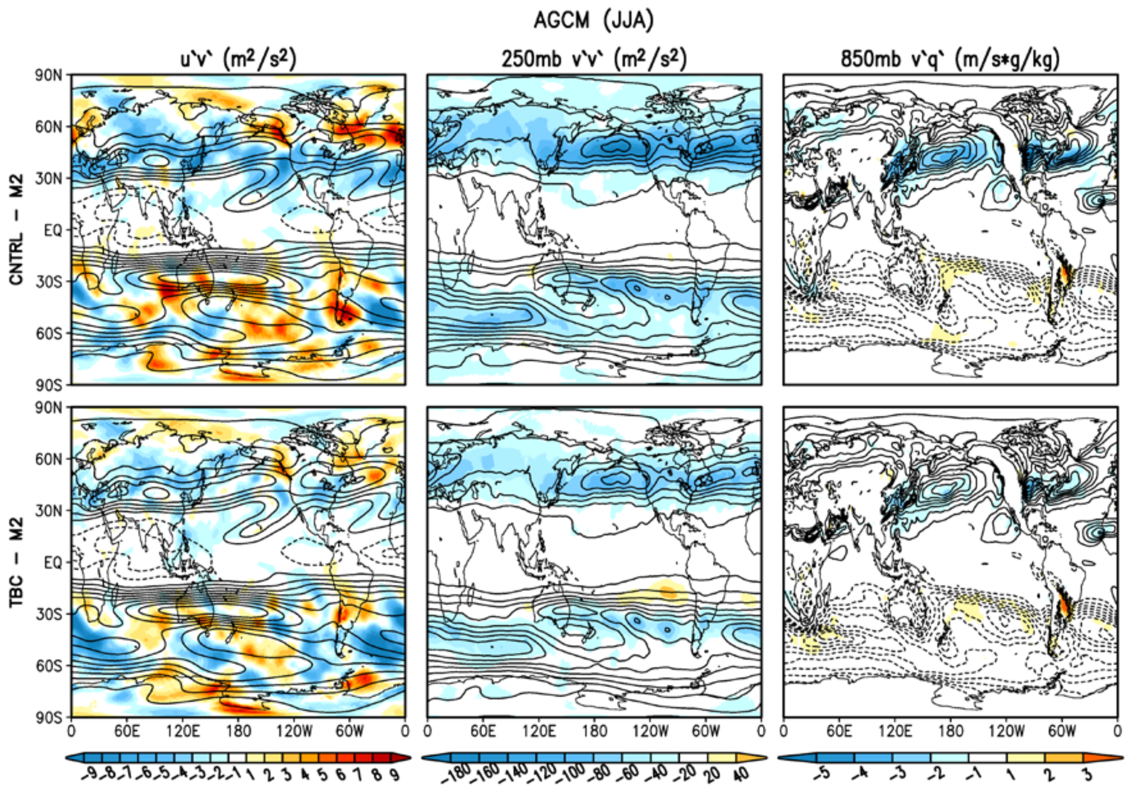

Figure 4: The vertically-integrated momentum flux by the transients ($(m/s)^2$, left panels), the 250mb square of the transient component of the meridional wind ($(m/s)^2$, middle panels),and the 850mb moisture flux by the transients (g/kg m/s, right panels). The shading indicates CNTRL-A – MERRA-2 in the top panels, and TBC-A – MERRA-2 in the bottom panels. In the left panels the contours are the 250mb climatological zonal wind from MERRA-2 (every 5 m/s). In the middle panels the contours indicate the long-term mean of the 250mb square of the transient component of the meridional wind from MERRA-2 (every 50 $(m/s)^2$). In the right panel the contours indicate the long-term mean of the 850mb moisture flux by the transients from MERRA-2 (every 1 g/kg m/s).

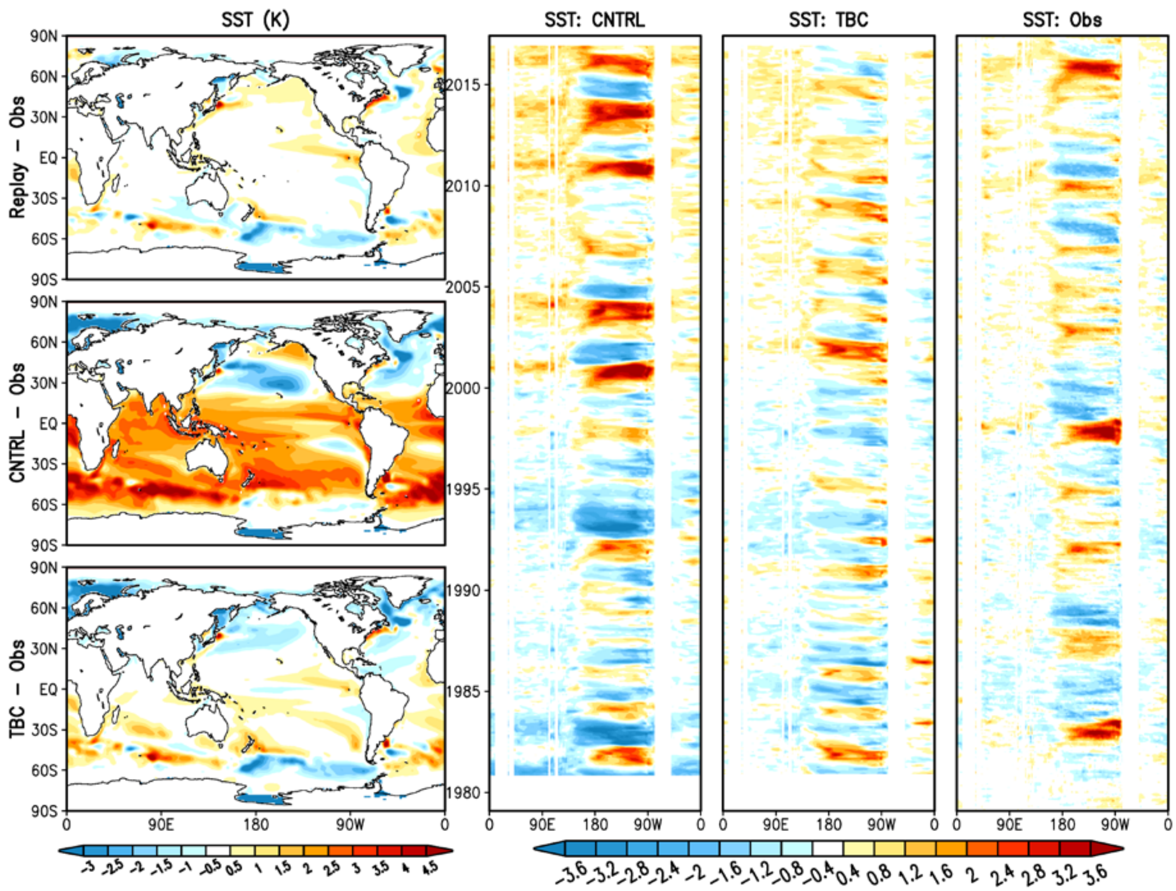

Figure 5: Left panels: the long term mean SST bias with respect to observations (ERSST.v4: Huang et al. 2015). Results are shown for REPLAY-C (top panel), CNTRL-C (middle panels), and TBC-C (bottom panel). Right panels: evolution of the monthly mean equatorial SST anomalies (2°S-2°N) from 1980-2016, for CNTRL-C, TBC-C, and the observations. Units: °K. Units: °K.   All fields are averaged over the years 1981-2016.

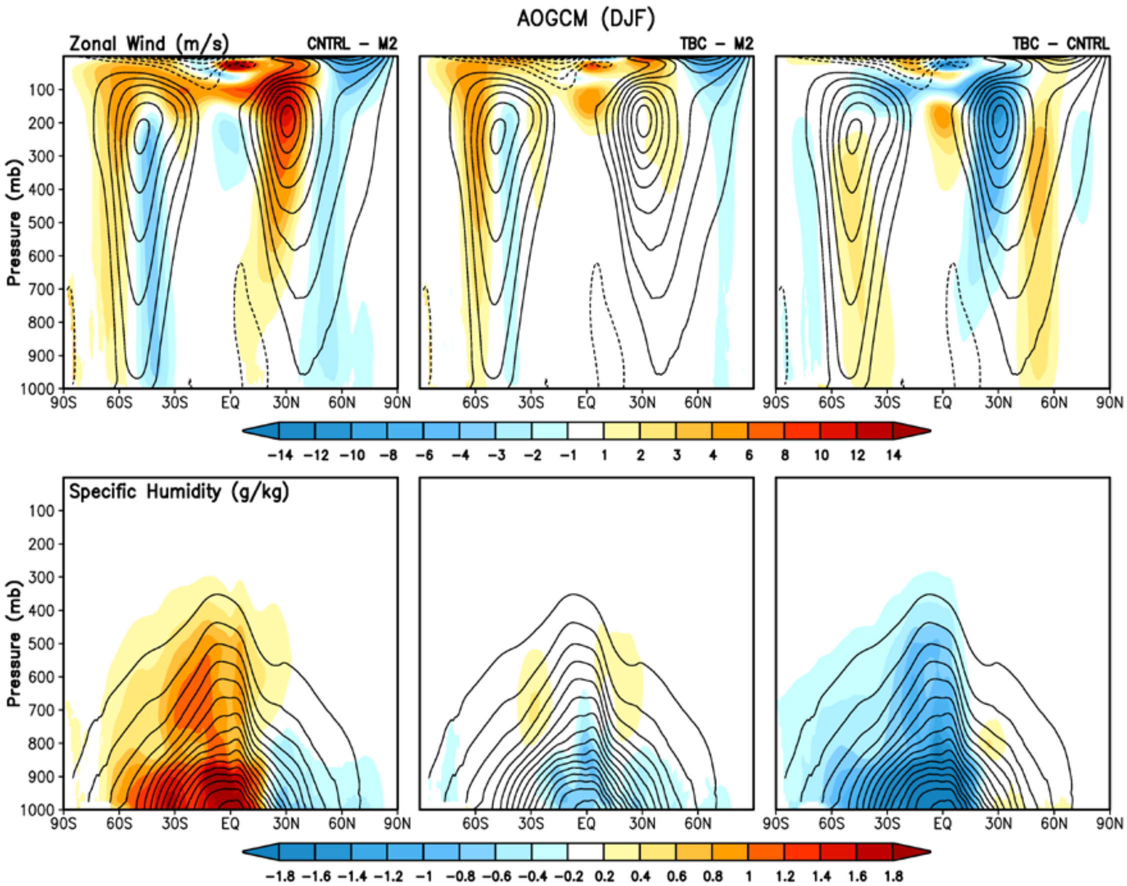

Figure 6: The zonal mean u-wind (top panels, m/s) and specific humidity (bottom panels, g/kg). Left panels: the shading indicates CNTRL-C – MERRA-2 with the climatological MERRA-2 wind fields contoured every 5 m/s in the top panels, and the MERRA-2 climatological specific humidity contoured every 1 g/kg in the bottom panels. Middle panels are the same as the left two panels, except for TBC-C - MERRA-2. Right panels the same as the left two panels, except for TBC-C – CNTRL-C. All fields are averaged for DJF over the years 1980-2015.

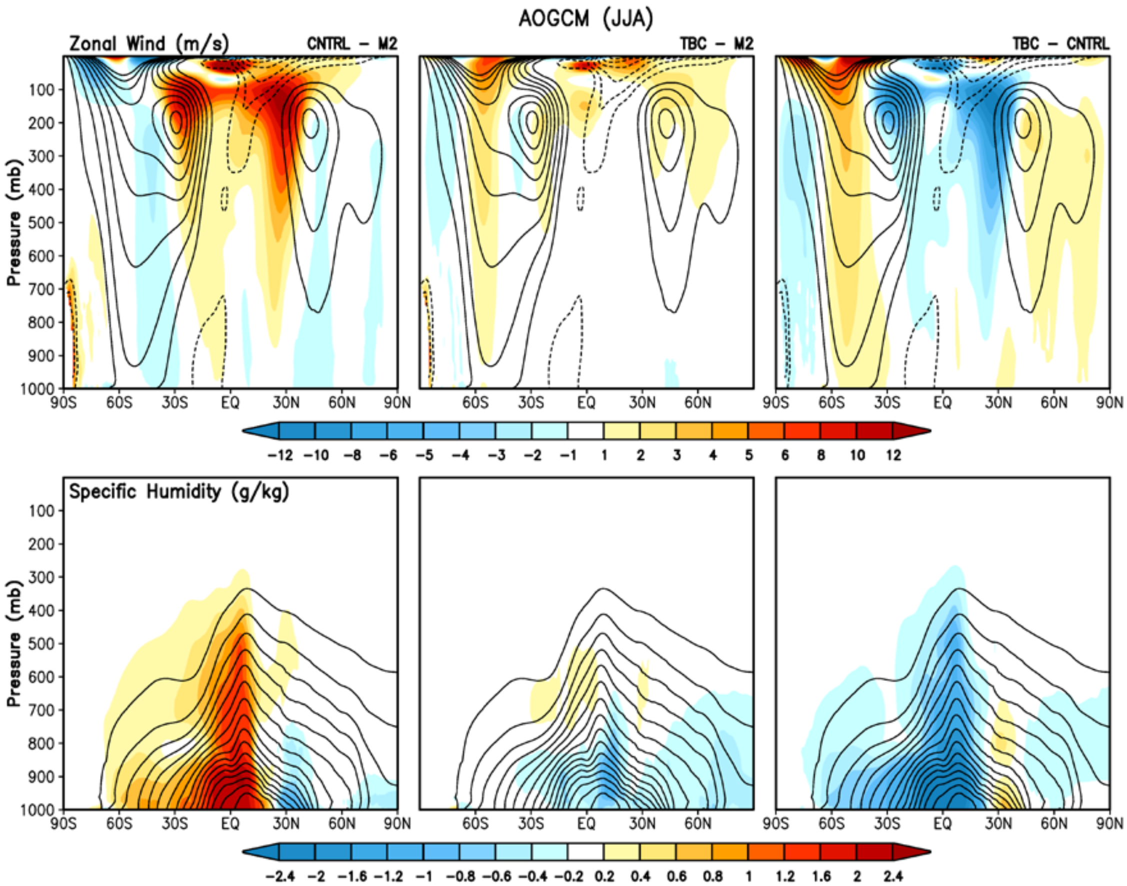

Figure 7: Same as Fig. 6, except for JJA.

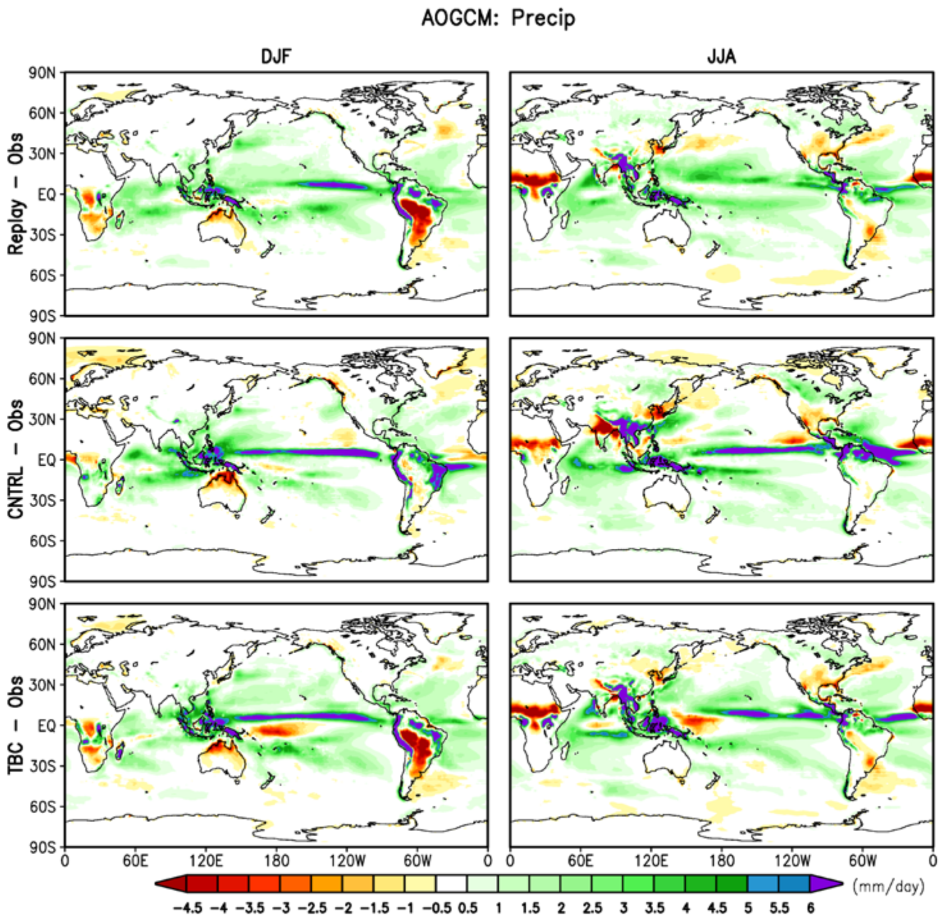

Figure 8: The precipitation biases (mm/day) with respect to MERRA-2 observationally corrected precipitation for DJF (left panels) and JJA (right panels). Results are shown for the replay run (top panel), the control (middle panels), and the TBC run (bottom panel). All fields are averaged over the years 1981-2016.

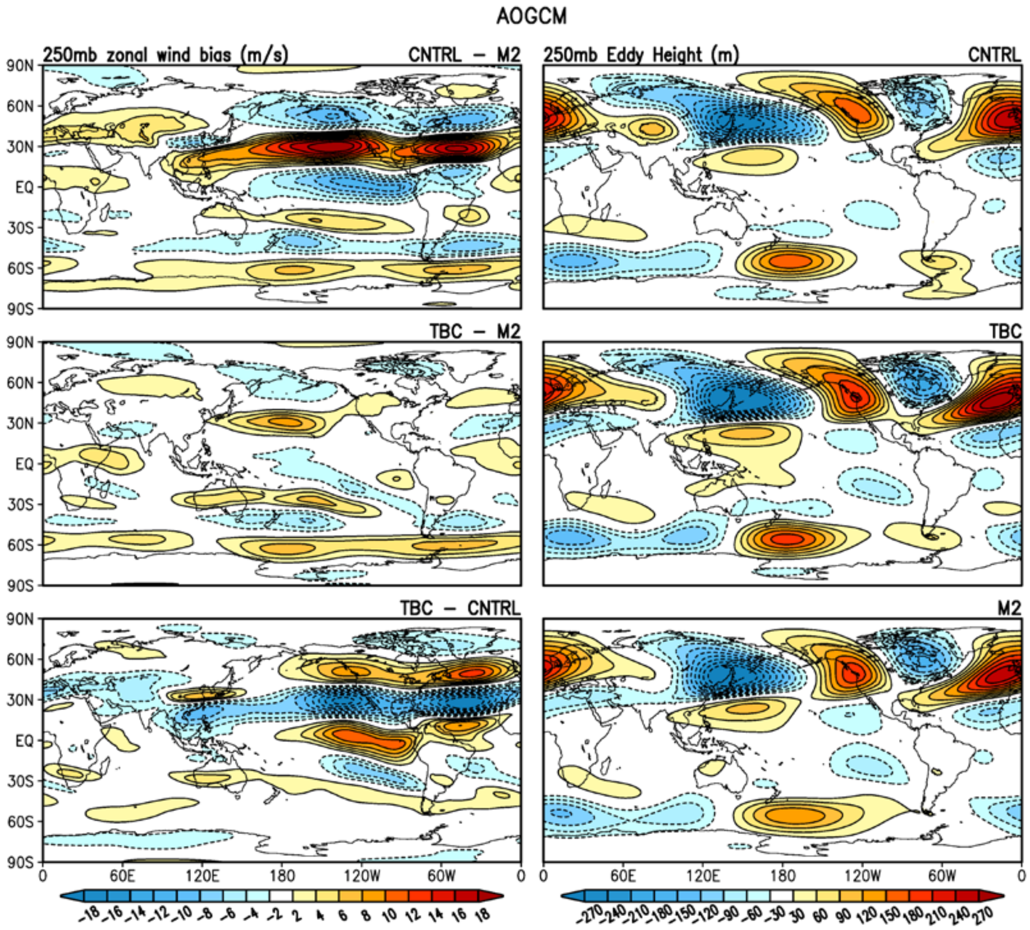

Figure 9: Left panels: the DJF zonal wind biases with respect to MERRA-2 for CNTRL-C (top) and TBC-C (middle). The bottom panels show TBC-C – CNTRL-C. Units are m/s. Right panels: the DJF stationary waves (250mb height with the zonal mean removed) for CNTRL-C (top), TBC-C (middle), and MERRA-2 (bottom). Units are meters.

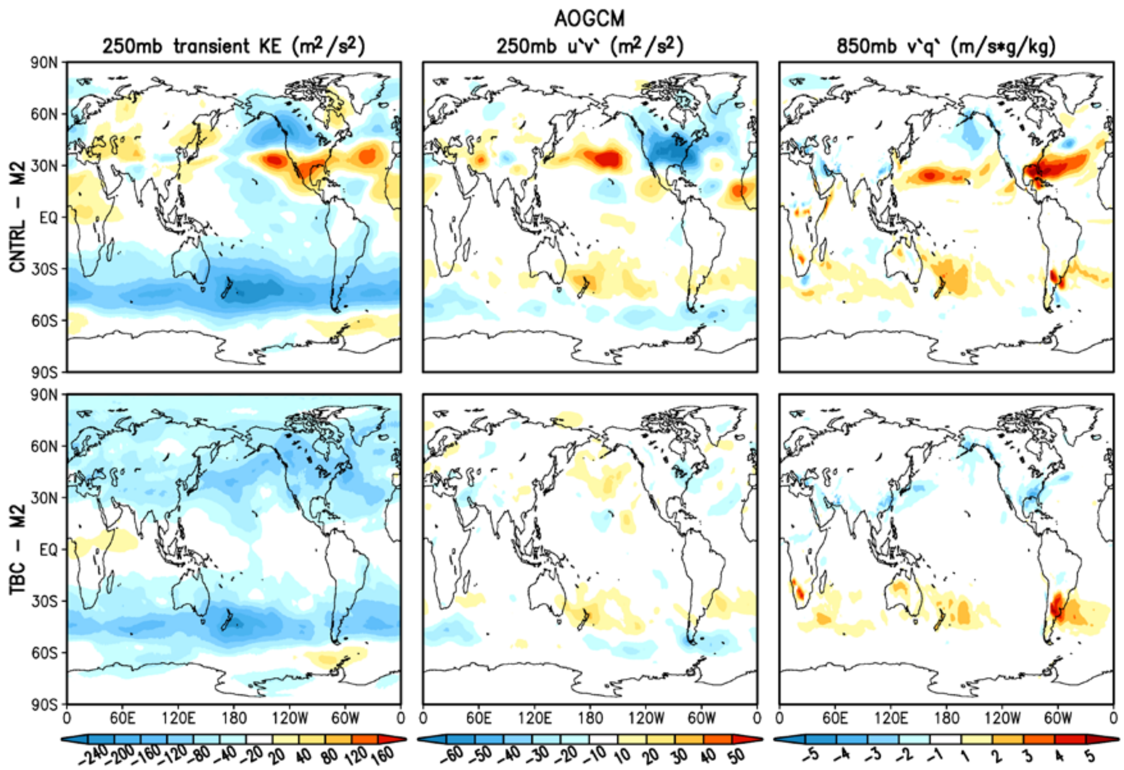

Figure 10: Left panels: CNTRL-C – MERRA-2 (top) and TBC-C – MERRA-2 (bottom) of the 250mb kinetic energy associated with the transient component of the winds (m/s)^2. Middle panels: same as left panels, but for the 250mb zonal momentum flux by the transients (m/s)^2. Right panels: same as left panels but for the 850mb moisture flux by the transients (g/kg m/s). All fields are averaged for DJF over the years 1981-2016.

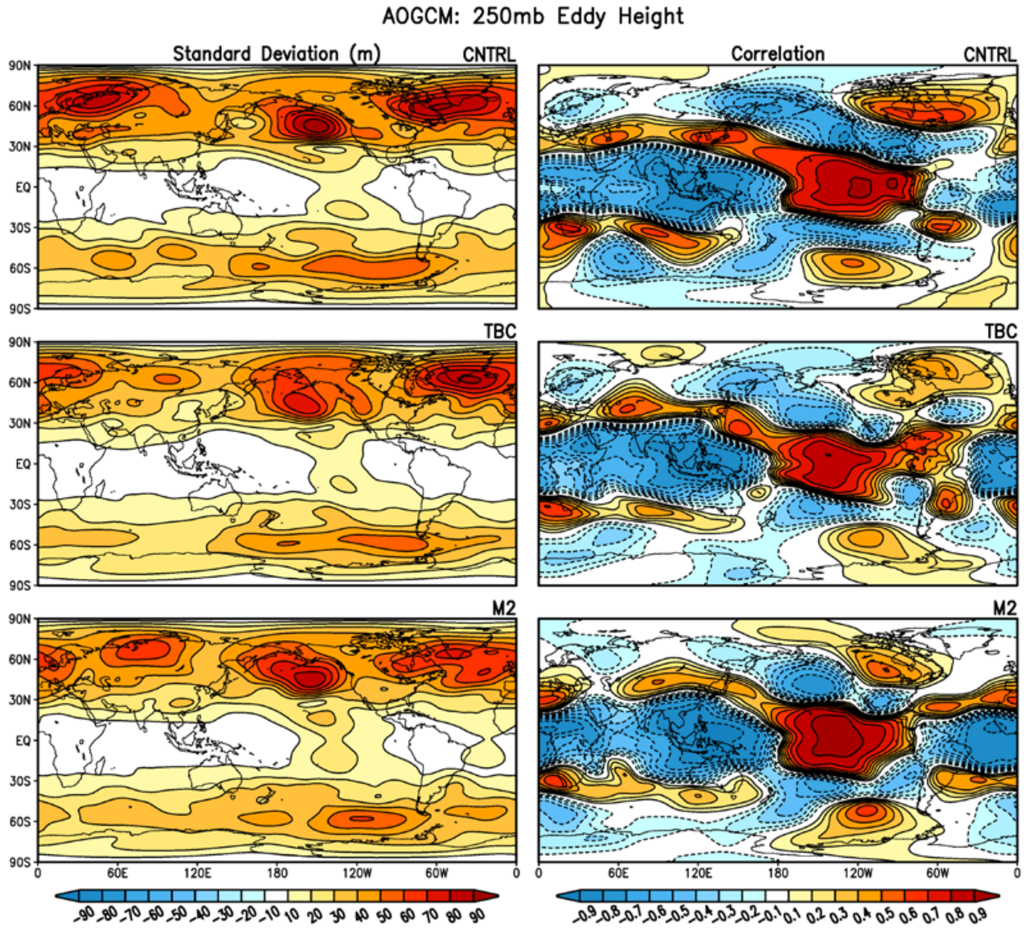

Figure 11: Left panels: the standard deviation (1981-2016) of the DJF mean stationary waves (250mb height with the zonal mean removed) for CNTRL-C (top), TBC-C (middle), and MERRA-2 (bottom). Units: m. Right panels: the correlations between the DJF mean Nino3.4 index and the 250mb eddy height field for the CNTRL-C (top), TBC-C (middle) and MERRA-2 (bottom).

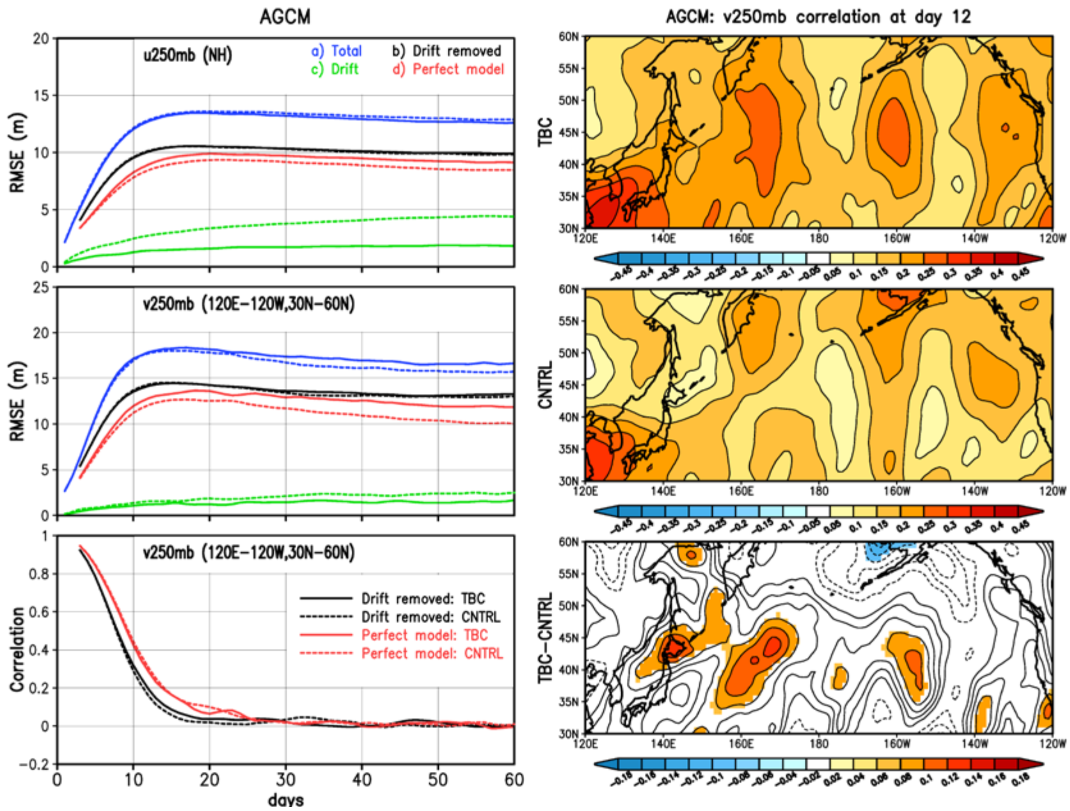

Figure 12: left panels: The RMSE decomposed according to (1) in the text for the daily 250mb u-wind for the NH (upper panel), and for the daily 250mb v-wind over the region (120E-120W, 30N-60N) in the middle panel. In the bottom panel, the black curves show the v250mb correlations with MERRA-2 (while the red curves are the correlations for a perfect model), for the same region (120E-120W, 30N-60N). The dashed lines are for the control hindcasts, and the solid lines are for the TBC hindcasts. Units for RMSE are m/s. Abscissa indicates days. Right panels: The v250mb correlations at 12 day lead for TBC-A, CNTRL-A and the differences. Shading of the differences indicates a significance level of 0.10 based on a Fisher's z-transform. Results are based on predictions initialized every day from May 1 to June 30 in 1988, 1998, 2000-2015. Five-member ensemble means are computed from lags -2,-1,0, 1, 2 days. See text for details.

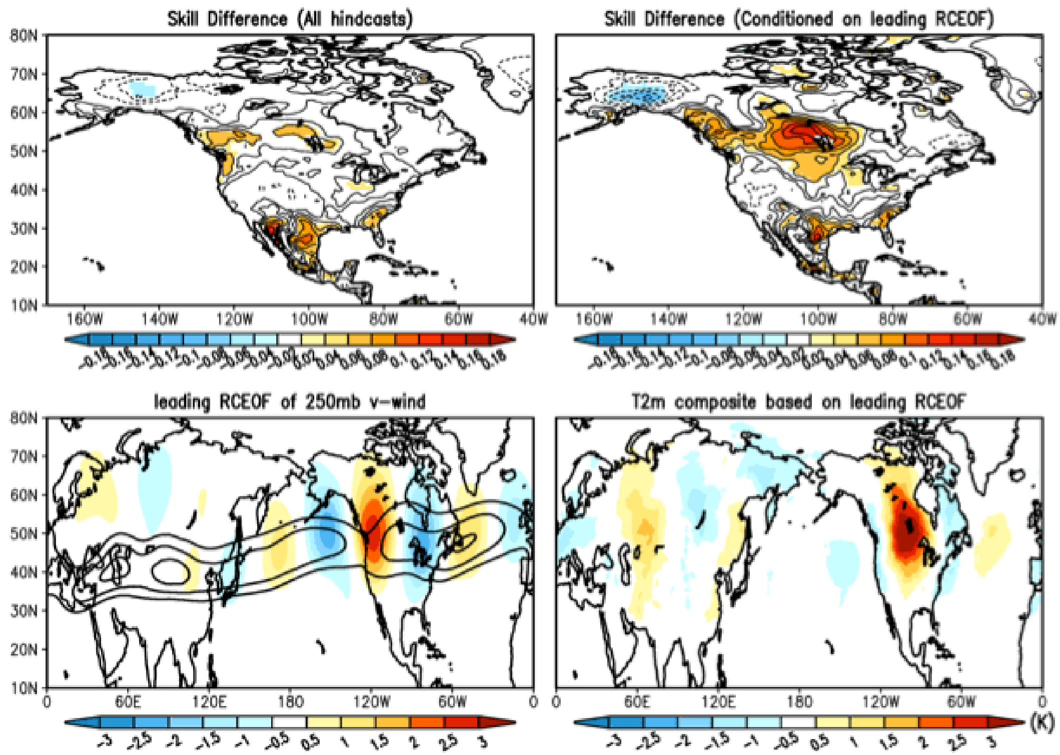

Figure 13: Top left panel: Differences in skill (correlation between the hindcasts and MERRA-2) at day 10 day between TBC-A and CNTRL-A based on all the hindcasts. Shading indicates a significance level of 0.10 based on a Fisher's z-transform. Top right panel: same as top left, except for only those hindcasts when the leading Rossby wave has an amplitude greater than 1 standard deviation in the initial conditions. Bottom left panel: The leading rotated complex empirical orthogonal function (RCEOF) of the NH (10N-80N) daily (filtered with a 11-day running mean) 250mb meridional wind anomalies during MJJA computed from MERRA-2 for the period 1980-2017 (see Chang et al. 2001 for details of the RCEOF calculation). The contours (15, 20 and 25 m/s) are the long-term mean MJJA 250mb zonal wind based on MERRA-2. The phase of the RCEOF plotted here is chosen to highlight that phase during which the wave has the greatest impact on North America (bottom right). The values of the RCEOF (m/s) and T2m (°C) correspond to composites based on those times when the associated rotated complex principle component (RCPC) exceeded 1 standard deviation.

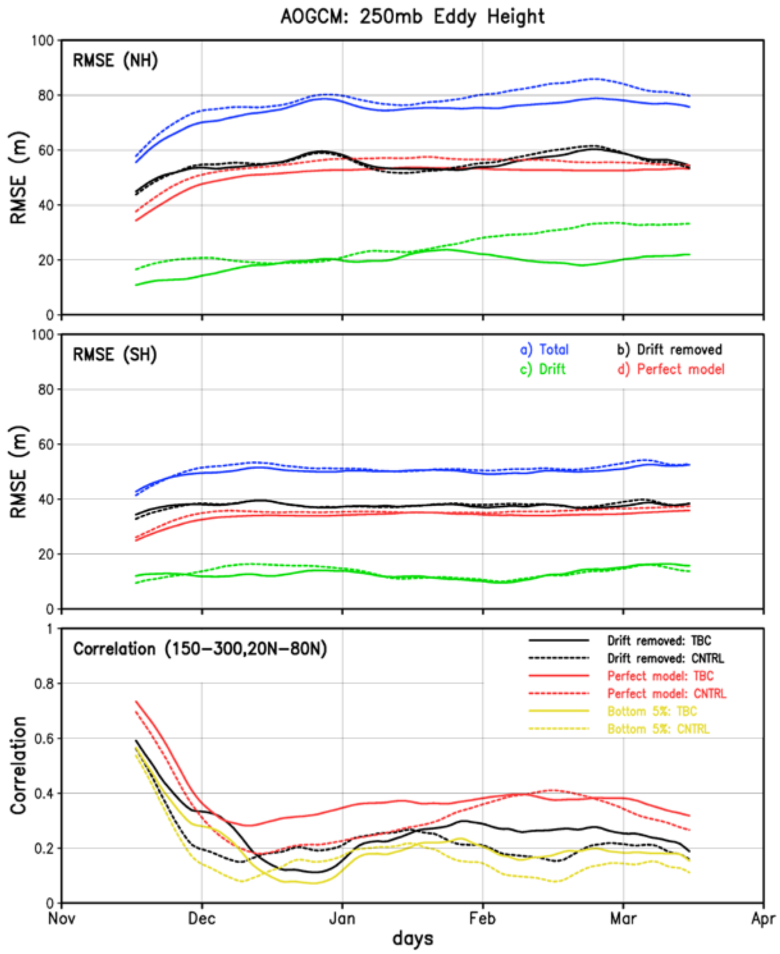

Figure 14: The RMSE decomposed according to (1) in the text for the 250mb eddy height for the NH (top) and SH (middle). The dashed lines are for the CNTRl-C hindcasts, and the solid lines are for the TBC-C hindcasts. Units are meters. Bottom panels show the PNA region 250mb eddy height correlations with MERRA-2 (black lines) and for a perfect model (red lines). The yellow lines, which are the bottom 5% of the correlations with MERRA-2 obtained from all combinations of removing 5 years from the 31 years of data (total of 169911), give an indication of the robustness of the 250mb eddy height correlations with MERRA-2. The daily fields have a 31-day running mean applied to remove weather and other sub-monthly noise. Results are based on 10 ensemble members initialized November 1 of 1985-2015.

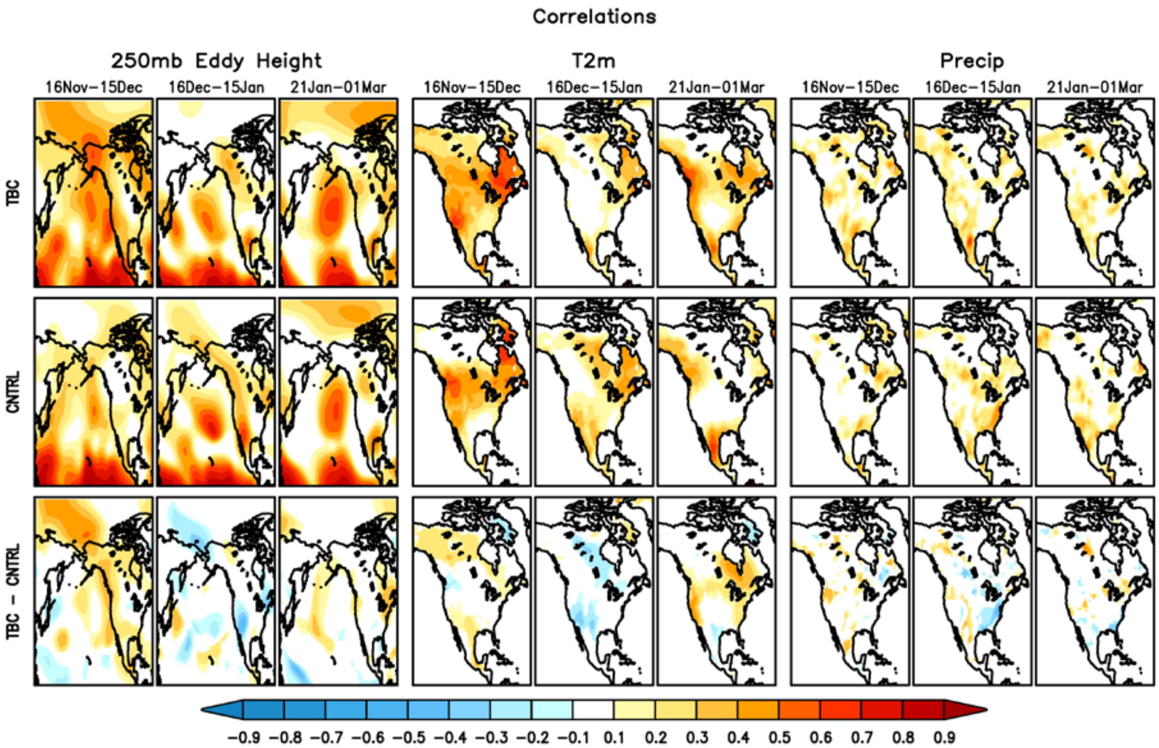

Figure 15: Maps of the correlations (only where values are greater than 0) between the hindcast ensemble mean and observations for 250mb eddy height (left 9 panels), T2m over North America (middle 9 panels), and precipitation over North America (right 9 panels). Results are based on 10 ensemble members initialized November 1 of the years 1985-2015. Top panels are for the TBC-C hindcasts, and middle panels are for the CNTRL-C hindcasts. The bottom panels are the differences in the correlations between the TBC-C and CNTRL-C hindcasts. Shading indicates a significance level of 0.10 based on a Fisher's z-transform. Results are shown for averages over the following time periods: 16Nov-15Dec, 16Dec-15Jan, and 21Jan – 01Mar.

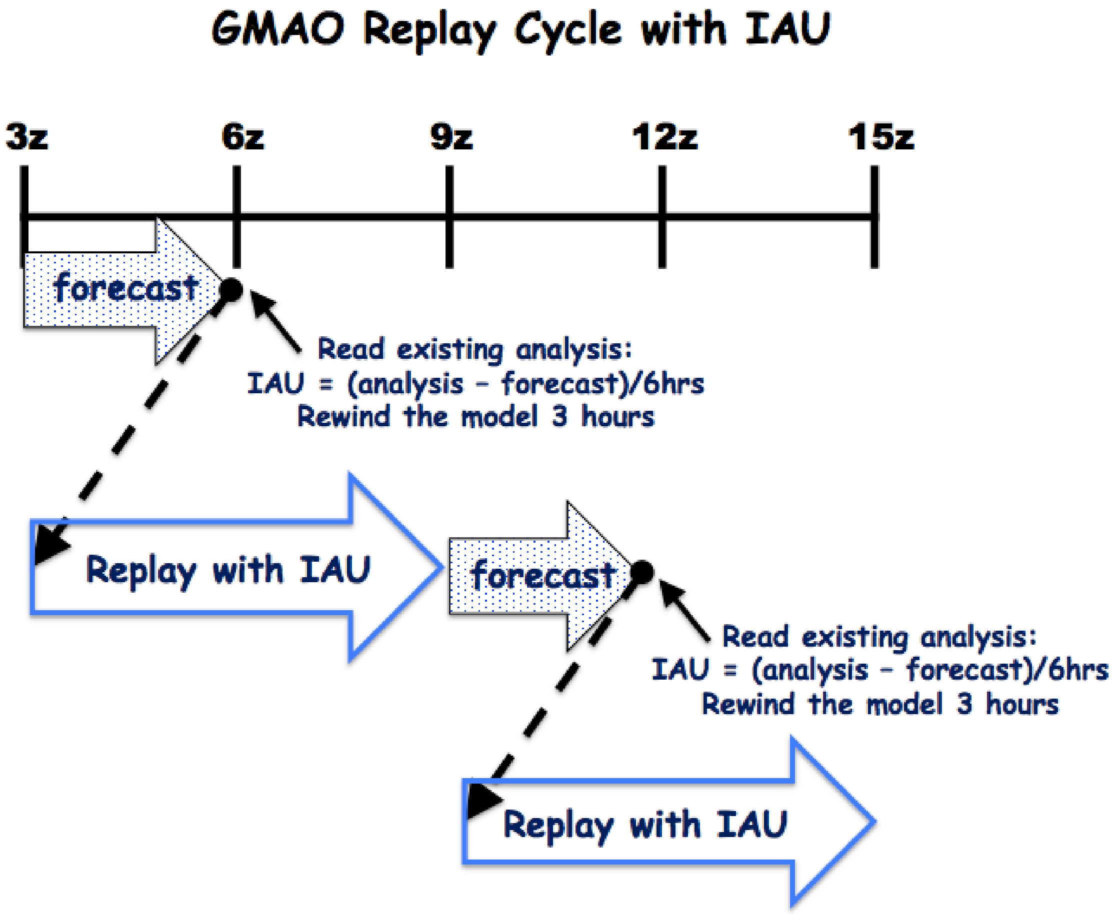

Figure A1: Schematic of the replay procedure used by the GMAO.