Improving seasonal forecasting through tropical ocean bias corrections

RMetS
Royal Meteorological Society

Improving seasonal forecasting through tropical ocean bias corrections

David P. Mulholland,[a]* Keith Haines[b]* and Magdalena Alonso Balmaseda[c]

[a]*Department of Meteorology, University of Reading, UK*
[b]*Department of Meteorology and National Centre for Earth Observation, University of Reading, UK*
[c]*European Centre for Medium-range Weather Forecasts, Reading, UK*

*Correspondence to: D. Mulholland or K. Haines, Department of Meteorology, University of Reading, Earley Gate, PO Box 243, Reading RG6 6BB, UK. E-mail: d.p.mulholland@reading.ac.uk; k.haines@reading.ac.uk

Initialization shock is often discussed in the context of coupled atmosphere–ocean forecasting, but its detection has remained elusive. In this article, the presence of initialization shock in seasonal forecasts is clearly identified in the variability of the tropical thermocline. The specific source of shock studied here is the use of a bias correction procedure to account for errors in equatorial wind stress forcing during ocean initialization. It is shown that the abrupt removal of the bias correction at the beginning of the forecast leads to rapid adjustments in the upper ocean, creating a shock that remains in the system for at least 3 months. By contrast, gradual removal of the correction term, over 20 days, greatly reduces the initialization shock. Evidence is presented of substantial increases in sea surface temperature (SST) seasonal forecast skill, at around 3–7 months' lead time, when the gradual removal approach is used. Gains in skill of up to 0.05, as measured by the anomaly correlation coefficient for SST in the Niño-4 region, are found, using a modest hindcast set covering four seasonal start dates. The results show that improvements in coupled initialization aimed at reducing shocks may considerably benefit seasonal forecasting.

Key Words: seasonal forecasting; climate prediction; ENSO; El Niño; data assimilation; ocean dynamics; shocks

Received 07 March 2016;; Revised 31 May 2016; Accepted 15 June 2016; Published online in Wiley Online Library 4 October 2016

1. Introduction

The initialization of coupled atmosphere–ocean dynamical models for forecasting on seasonal time-scales (∼3–9 months) currently poses a challenge to forecasters (Weisheimer *et al.*, 2009; Barnston *et al.*, 2010; Molteni *et al.*, 2011; MacLachlan *et al.*, 2014; Saha *et al.*, 2014). Observational information must be incorporated into both components of the models, particularly the ocean component, in order to realise skilful forecasts, but the optimum strategy for achieving this in the presence of biases that exist in all numerical climate models is not yet known (Balmaseda and Anderson, 2009; Magnusson *et al.*, 2013; Smith *et al.*, 2013; Thoma *et al.*, 2015). One requirement of a successful model initialization procedure is that it avoids the generation of initialization shocks (Rahmstorf, 1995; Chen *et al.*, 1997; Zhang *et al.*, 2007; Zhang, 2011; Mulholland *et al.*, 2015; Smith *et al.*, 2015), due to inconsistencies or imbalances in the initial model state, at the beginning of the forecast, a problem which stretches back to the numerical weather predictions of Richardson (1922).

In coupled models, initialization shocks may be particularly prevalent in the Tropics, where the ocean and atmosphere are strongly coupled. In particular, close to the Equator, any imbalances between zonal wind stress and the zonal pressure gradients in the upper ocean can lead to the generation of subsurface ocean waves which propagate along the thermocline, later affecting sea surface temperatures (SST) in the eastern part of the ocean basins where the thermocline lies close to the surface (Harrison and Giese, 1988; Vecchi and Harrison, 2000). Via reflections at the basin boundaries, such signals can then remain in the ocean for O(1 year), and potentially longer via coupled feedbacks (Fedorov, 2002). Further, through atmospheric teleconnections, signals in tropical SST can exert a global impact (e.g. Trenberth *et al.*, 1998; Mason and Goddard, 2001; Mathieu *et al.*, 2004; Sanchez-Gomez *et al.*, 2016). Zonal imbalances can easily arise when tropical mooring data, such as from the Tropical Ocean–Atmosphere (TAO) array (Hayes *et al.*, 1991) are assimilated, leading to changes in ocean zonal pressure gradients. Although this problem is well known (Bell *et al.*, 2004), it is still unclear whether this results in degradation in seasonal forecast skill, via nonlinear interactions which cannot be corrected by linear post-processing drift correction (Stockdale, 1997).

Fundamentally, this problem arises because of the relatively large uncertainties associated with atmospheric reanalysis wind speeds in the Tropics (Kent *et al.*, 2013) and errors in the parametrization of the vertical propagation of turbulent wind

stress through the upper ocean. These result in biases in thermocline structure and in SST along the Equator, in both ocean models forced with reanalysed winds, and free-running atmosphere–ocean coupled models. Data assimilation attempts to correct these biases, using subsurface density information, but this creates large spurious circulations in the ocean analysis when the model effectively 'rejects' these corrections to the thermocline, due to the imbalance between zonal pressure gradient and wind stress forcing, leading to large adjustments at each analysis time step.

To preserve dynamical balance, and avoid these circulations, the 'pressure correction' bias scheme of Bell *et al.* (2004) was developed, and is now commonly used in operational systems, e.g. at the Met Office (Blockley *et al.*, 2014) and ECMWF (Balmaseda *et al.*, 2013). The pressure correction scheme modifies the pressure gradient forces in the tropical upper ocean, in order to allow the correct average thermocline structure (and hence realistic SST) to exist in the presence of erroneous wind stresses and/or vertical ocean mixing. The correction term may have a slowly varying (seasonal cycle) component, using a monthly climatology from a previous run of the ocean reanalysis, along with a higher frequency component, calculated from errors occurring on a time-scale of a few days (Balmaseda *et al.*, 2007; Mogensen *et al.*, 2012).

However, in forecast mode no ocean observations are available, so the correction scheme is usually switched off, since ocean wind stress errors develop rapidly in the free-running coupled model, making the pre-calculated pressure correction term inappropriate. The model ocean must then adjust to establish a new balance between the zonal wind stress forcing (which it now 'sees' fully for the first time) and its density structure. This is the initialization shock which we will address in this article, and it occurs even before the free-running coupled model winds deviate from the winds used to force the ocean analysis, due to the instantaneous loss of the bias correction term. Forecast model winds will also drift from the truth, due to the existence of systematic biases, and this quickly dominates overall errors. However, the shock due to the removal of the bias correction term may interact nonlinearly with other drifts, and with the evolving forecast state, complicating the task of post-process forecast drift correction.

This choice of initialization methods is one example of a wider issue of seeking to find an optimum balance between creating accurate initial conditions close to the observed state, and allowing the model to remain consistent with its own, biased climatology, in order to avoid rapid adjustments in forecast mode when observational information is no longer available. It is similar to the choice between 'full-field' and 'anomaly' initialization (Pierce *et al.*, 2004; Smith *et al.*, 2007), in which anomalies from an observed climatology are assimilated into the model's own biased climatology. Anomaly initialization avoids large forecast drifts at the expense of a realistic mean state, which can adversely affect the anomaly forecasts. It is not yet clear whether or in what circumstances the anomaly initialization method might be superior to full-field assimilation (e.g. Meehl *et al.*, 2009; Magnusson *et al.*, 2013; Smith *et al.*, 2013; Polkova *et al.*, 2014).

In this article we investigate the impact of the use, and subsequent removal, of the pressure correction term within the seasonal forecasting system at the European Centre for Medium-range Weather Forecasts (ECMWF), and evaluate the performance of alternative initialization methods which avoid the instantaneous removal of the bias correction term. We test our hypothesis that avoiding or minimizing initialization shocks in the tropical oceans can lead to improved forecast skill at several months' lead time and beyond. The model used and the forecast sets performed are described in section 2. The results of the experiments are presented in section 3. Interpretations of the results and issues regarding the use of the various methods are discussed in section 4. Finally, the key results are summarised in section 5.

2. Model and experiments

2.1. Forecast system and ensemble composition

Four sets of forecasts were carried out using the ECMWF coupled forecasting system, consisting of the Integrated Forecast System (IFS) cycle 40R3, at a spectral resolution of T255 with 91 vertical levels, as the atmospheric component, and the Nucleus for European Modelling of the Ocean (NEMO) model v3.4, with a horizontal resolution of 1° in midlatitudes, increasing in the meridional direction to 0.3° at the Equator, and 42 vertical levels, as the ocean component, and a time step of 45 min. All forecasts ran for 7 months using ensembles of five members, initialized from a single ocean reanalysis, with perturbations applied in the initial SST and atmospheric states to generate the ensembles. The forecast sets included 16 start dates covering four seasons (four dates in each of February, May, August and November) in a range of years spanning 1980–2009 (chosen according to the availability of reanalysis initial conditions on the first day of the month, and including both El Niño and La Niña years). All sets used atmospheric initial conditions from ERA-Interim (Dee *et al.*, 2011), which was run using an older version (cycle 31R2) of IFS.

2.2. Forecast sets

The first set, named OP, had its ocean component initialized from the Ocean Reanalysis System (ORAS4; Balmaseda *et al.*, 2013), which used the pressure correction scheme of Bell *et al.* (2004) during its production. The pressure correction scheme as implemented in ORAS4 adds a zonal pressure gradient term, made up of climatological and 'online' components, to the ocean model momentum equations. The climatological component is derived from temperature and salinity biases computed, relative to ocean subsurface profile observations, from a previous run of the analysis which used relaxation to climatology instead of bias correction, averaged over the period 2000–2008. This is stored as a monthly seasonal cycle, and interpolated to the appropriate forecast date. The online term is updated in each analysis cycle, and is generally smaller than the climatological term (Balmaseda *et al.*, 2013). This correction scheme was then switched off at the beginning of the OP forecasts. This is the same approach as used in ECMWF's operational forecasting System 4 (S4) (Molteni *et al.*, 2011) and, as such, OP is taken as the baseline over which improvements in forecast skill are sought.

The second set, NOBC, was initialized from a different reanalysis, ORAS_nobc, that was created in a manner identical to ORAS4 but without using the pressure correction scheme during the analysis (Balmaseda *et al.*, 2013). Again, no bias correction was used during the forecasts.

The other two forecast sets are modifications of OP, with the equatorial bias correction now also applied during the forecasts. Only the climatological bias component was used during the forecasts, but this is the dominant contributor to the overall correction in ORAS4. Therefore, the only change in pressure gradient forces that occurs at the start of the coupled forecast is the loss of the small high-frequency bias correction component.

In set PERS, the bias correction term was applied at full strength throughout the 7-month forecasts (note that, although this set is named for a 'persisted' correction term, the term does in fact vary in time, as it is a monthly climatology). However, the climatological bias correction term was calculated from an uncoupled analysis using the ocean model forced by reanalysis winds, so may not remain useful during the seasonal forecasts, when substantial model drifts occur. Mean drifts will be altered by the continued application of the correction term in PERS, which could have positive or negative effects on forecast skill in different regions, depending on the agreement in sign between the drift and the correction term.

The final set of forecasts, DAMP, also used the bias correction, but with a scaling factor which decreases linearly to zero over a time window of 20 days from the beginning of the forecast. This was chosen to match the drifts in near-surface zonal wind along the Equator, which occur over ∼ 10 days in the ECMWF system, while still ensuring that the correction term was not removed too rapidly. This aims to avoid prolonged use of a sub-optimal correction field while at the same time avoiding the initialization shock that is potentially present in OP due to the instantaneous removal of the field. Note that the methods PERS and DAMP require no future information, as they use only the climatological bias correction term, so are viable strategies for real-time forecasting.

To measure seasonal forecast skill, the root-mean-square error (RMSE) and anomaly correlation coefficient (ACC) were calculated for SST using NOAA's Extended Reconstruction SST dataset (ERSSTv4; Huang *et al.*, 2015) as the common reference for all forecast sets. The ACC and RMSE results presented here do not differ greatly if ORAS4 or ORAS4_nobc SST are used as references instead. ACC is insensitive to mean forecast drift, so ACC differences between forecast methods should be due to nonlinear interactions that cannot be removed through an *a posteriori* drift removal. The statistical significance of differences in ACC or RMSE time series between pairs of forecast sets was computed using a bootstrap sampling method over the 16-date forecast sets (Appendix). However, 'climatologies' for each forecast set are computed separately for each of the four seasons using only four dates, so values contain further uncertainty resulting from undersampling of climate variability in each season, and must still be viewed with caution.

2.3. *Effect of the bias correction term*

Differences in the initial conditions of OP and NOBC are shown in Figure 1 for ocean density and zonal velocity along the Equator, averaged over all 16 start months. Seasonal variations (not shown) are fairly small but not negligible, particularly in the Indian Ocean. In the equatorial Pacific, the effect of the pressure correction is to increase the density (reduce the temperature by up to 1 °C) and raise the thermocline by ∼ 2 m around 100–150°W, and to deepen the thermocline by 1–2 m close to the dateline and at the eastern boundary. The associated circulation response is upward and eastward at the depth of the thermocline at 120°W, and westward at the surface. A similar pattern is seen in the Atlantic basin, while the response in the Indian basin includes a slightly weaker shallowing of the thermocline at 40–60°E.

These differences make the initial conditions of OP (and PERS and DAMP) more consistent with observations, whereas NOBC is placed at a disadvantage by the neglect of bias correction in its initial conditions (Balmaseda *et al.*, 2013). However, this also implies that the initial surface and subsurface temperature distributions in OP cannot be sustained by the model with the bias correction field, so OP forecasts are expected to drift rapidly in the opposite direction to the fields shown in Figure 1. That is, for example, a downwelling adjustment should occur in the eastern Pacific thermocline depth at 100–150°W, and an upwelling adjustment should occur along the thermocline at around 180°E. Also, westward surface currents should decelerate in the central Pacific and eastern Atlantic oceans. These adjustments, or initialization shocks, will be superimposed on ocean drifts that are driven by the atmospheric model during the first few days of the forecasts; e.g. if near-surface winds strengthen over the central Pacific, this could reverse the surface zonal current response to be one of acceleration, and could enhance the upwelling adjustment occurring in the underlying ocean.

3. Results

3.1. *Thermocline response to initialization shock*

The ensemble mean time series of the 20 °C isotherm depth averaged in Niño-4 (160°E–150°W, 5°N–5°S) for the first 30 days of each forecast set are shown in Figure 2(a), alongside the reanalyses ORAS4 and ORAS4_nobc. A shallow (linear) drift of the forecasts relative to the reanalyses is clear, caused by model biases. Variability with period 4–5 days is apparent in all forecasts, and some variability on this time-scale can also be seen in the reanalyses. Waves at this high frequency are slightly stronger in OP relative to the other forecasts, and there is an additional small (less than 1 m) upward thermocline adjustment in the first 1–2 days, as was predicted in the previous section to occur at 170–180°E following the sudden removal of the bias correction field. The phase of the high-frequency wave in OP is also shifted by roughly half a cycle, while the other forecasts remain roughly in phase with the reanalyses out to 30 days' lead time. (The slight downward adjustment from the reanalyses seen on day 1 in all forecast sets is another form of shock, and occurs as a consequence of the neglect of surface currents in the wind stress applied to ORAS4 and ORAS4_nobc, possibly combined with atmospheric wind drifts within the first 24 h; section 3.2.)

The frequency spectra in Figure 2(b) show clearly the additional wave energy present in the OP forecasts compared to the others,

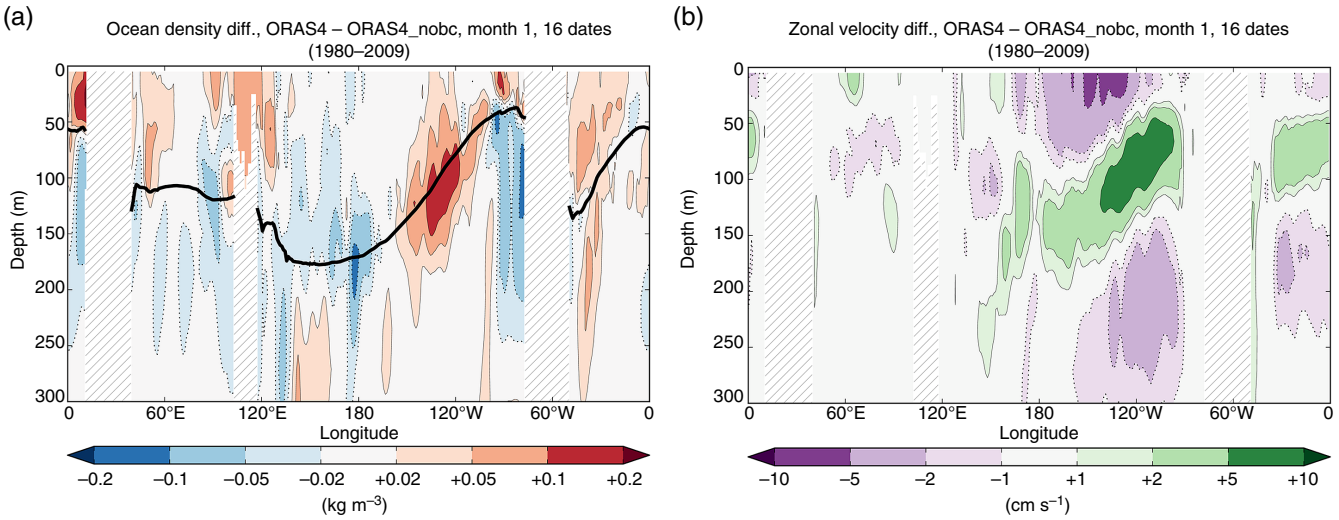

Figure 1. Differences between the ORAS4 and ORAS4_nobc reanalyses in (a) density (kg m^{-3}) and (b) eastward zonal velocity (cm s^{-1}) at 5°N–5°S, averaged over all 16 forecast start months. Solid (dotted) contours mark positive (negative) values. Land areas are hatched. In (a), the mean 20 °C isotherm in ORAS4_nobc is marked by a black line.

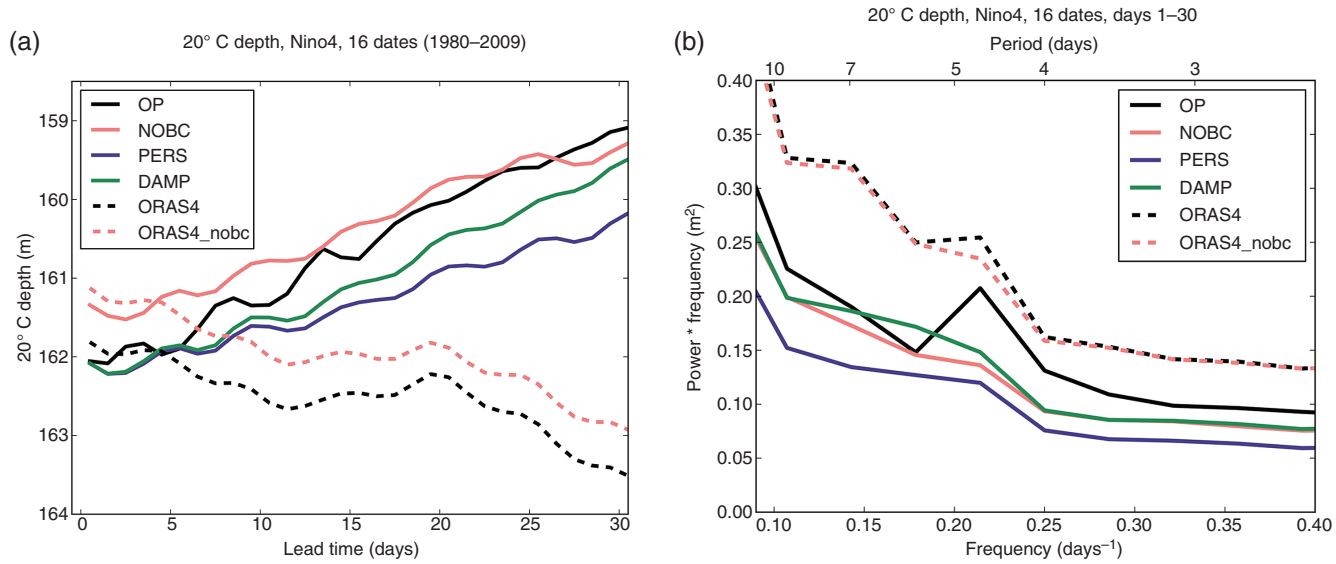

Figure 2. Niño-4 20 °C isotherm depth: (a) ensemble mean time series (m) and (b) ensemble mean frequency spectra (m^2), for the four experiments (solid) and two reanalyses (dashed). Values plotted in (a) and used to calculate (b) are daily means, starting at day 1.

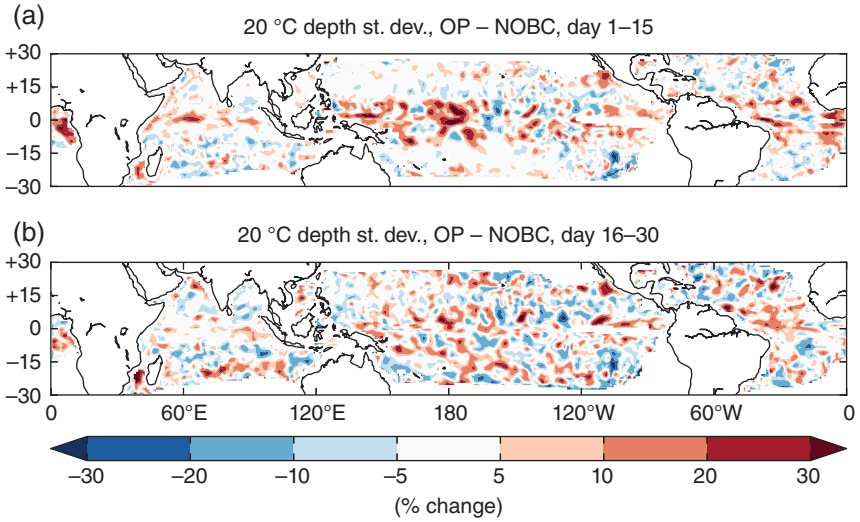

Figure 3. Percentage increase in the average temporal standard deviation of the 20 °C isotherm depth time series in OP relative to NOBC, for days (a) 1–15 and (b) 16–30. The diagnostic cannot be calculated beyond a latitude of ∼25°N/S, where the surface temperature falls below 20 °C.

with a clear peak in the spectrum at 0.20–0.25 d^{-1} (period 4–5 days). This enhancement in wave energy in OP can be regarded as the signature of an initialization shock that occurs following the instantaneous removal of the bias correction term. The spectra confirm that the additional thermocline variability is not present in PERS, and in DAMP it appears to have been almost entirely removed. There is greater power at a range of frequencies in the two reanalyses, due to the forcing of the thermocline by assimilated subsurface observational data, particularly where these data capture mesoscale eddy processes that are at the limit of the model's horizontal resolution (30 × 110 km at the Equator).

Figure 3(a) confirms that day 1–15 differences in 20 °C isotherm depth temporal variability between OP and NOBC are initially largely confined to the Tropics (15°N–15°S), where the pressure correction method was applied in ORAS4. This is further evidence that the differences between OP and NOBC are due to the sudden removal of the bias correction term at the beginning of the OP forecasts, rather than the different initial model states (Figure 1) *per se*. At days 16–30 (Figure 3(b)), differences in 20 °C isotherm depth variability have spread to higher latitudes, suggesting that the long-term effects of initialization shock are not limited to the region in which the bias correction term is applied.

Substantial initialization shocks are present in the western parts of all three tropical basins (Figure 3(a)), where the thermocline is deepest. Anomalies in thermocline depth tend to propagate eastward as equatorial Kelvin waves, so the subsurface shock in the west can later influence the eastern parts of the basins, adding to any shock experienced locally there. Further, any shock effects felt in tropical SST have the potential to affect the global atmosphere via teleconnections. Extratropical weather patterns have been noted to be particularly sensitive to SST anomalies in the western/central tropical Pacific area (Barsugli and Sardeshmukh, 2002).

The amplitude and persistence of the initialization shocks in OP, NOBC and DAMP can be seen in Figure 4, which shows averaged wavelet spectra (Torrence and Compo, 1998) for Niño-4 20 °C isotherm depth, for the first 60 days of the forecasts. Spectra were calculated for each of the 16 ensemble mean forecast series, and averaged for each forecast set and for ORAS4. The ORAS4 spectrum shows stronger power in the 4–8 day band at 10–20 days' forecast lead time, despite being averaged over multiple start dates, indicating that edge effects are present in the calculation even beyond the marked 'cone of influence'. Nevertheless, the initialization shock can be seen primarily as the enhancement of power at 4–8 days' period at 10–50 days' lead time in OP (Figure 4(a)) relative to the other forecasts (Figure 4(b, c)). In DAMP (Figure 4(c)), the power at 4–8 days' period is weaker than in OP by a similar amount to NOBC over the first 50 days, confirming that the initialization shock has been largely avoided

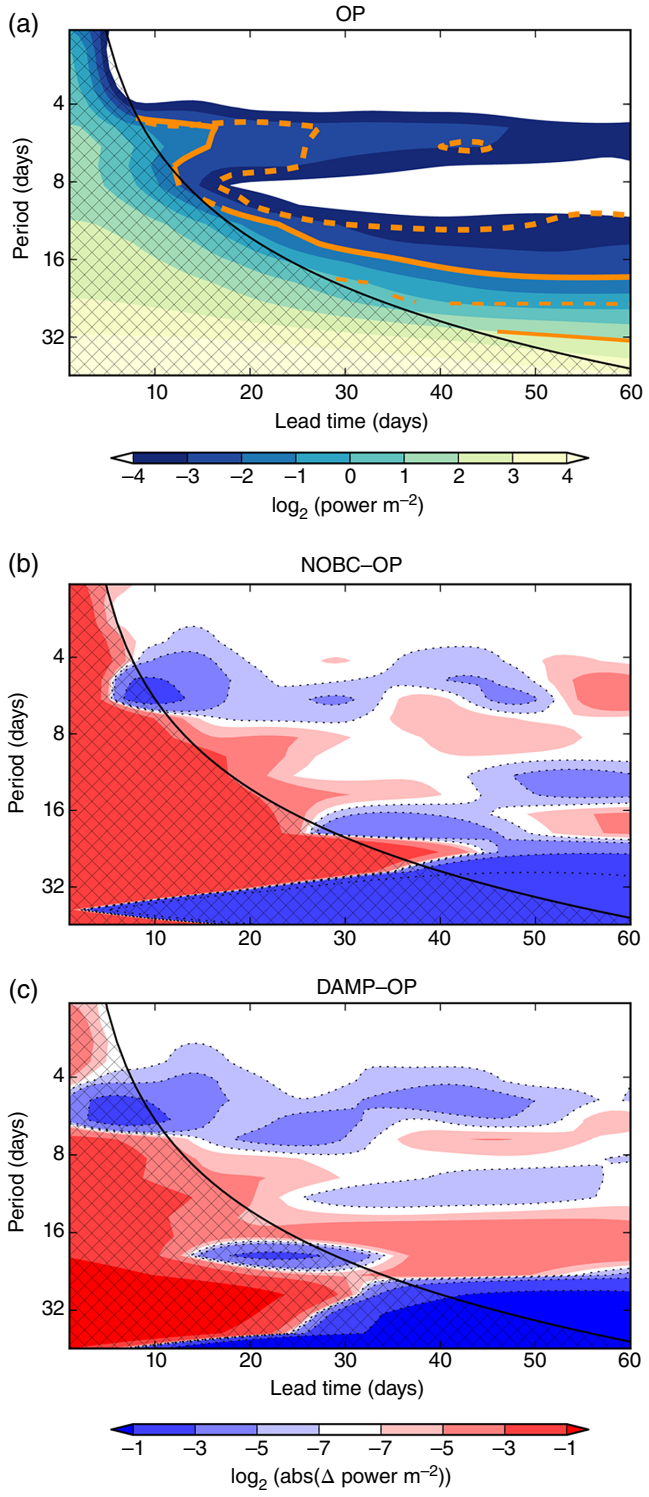

Figure 4. (a) Wavelet power spectra (m^2, contoured on a log scale) for Niño-4 20 °C isotherm depth, computed using time series at daily frequency for the full 210 day duration of the forecasts, and averaged over all start dates, for the first 60 days of OP (shading, and thin (thick) solid contours at 2^2 (2^{-2})). Thin (thick) dashed contours mark values of 2^2 (2^{-2}) from ORAS4. (b, c) Average difference in wavelet power spectra for NOBC−OP and DAMP−OP, respectively (shading; dotted contours for negative values). Contour levels are on a log scale, with negative (positive) values where NOBC/DAMP show lower (greater) power than OP. Values in the hatched areas, at lead times less than approximately $\sqrt{2}$ times the period, are outside the 'cone of influence' and are contaminated by edge effects.

in DAMP, by virtue of the more gradual removal of the correction field. The strong appearance of the initialization shock in the 4–8 days band is likely due to the excitation of a gravity wave mode, perhaps with some numerical interaction with the model spatial resolution, or the surface flux or SST forcing time-scales imposed in ORAS4.

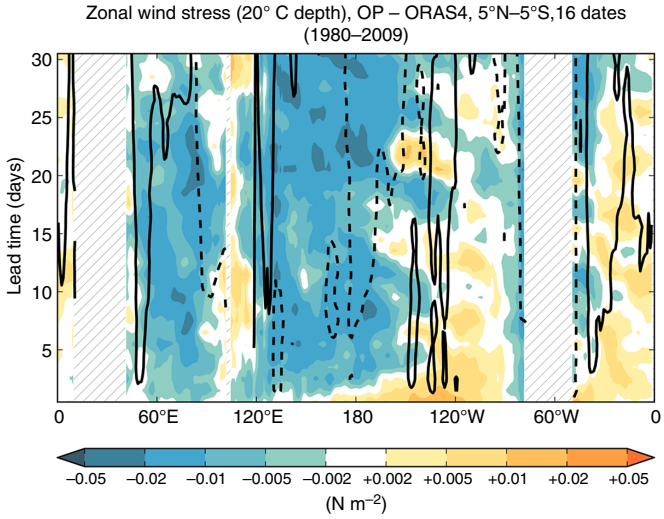

Figure 5. Mean drift in surface zonal wind stress (shading, N m^{-2}) and 20 °C isotherm depth (contours at +2 (solid) and −2 m (dashed)) in OP relative to ORAS4 in the first month, averaged over 5°N–5°S. Values are non-zero on the first day at some longitudes due to the calculation of surface stresses relative to surface currents in the coupled model, but relative to a static surface in ORAS4, as well as drifts in atmospheric winds within the first day. Land areas are hatched.

There are also suggestions of reduced power at longer periods as the forecasts progress, at lead times of 40–60 days (Figure 4(b, c)) and longer (not shown). This is a possible pathway for initialization shock to affect seasonal forecast skill (sections 3.3 and 4.1).

3.2. *Interaction with model drifts*

We now look at the rate of development of model errors, since these will also affect forecast skill. Figure 5 shows the forecast drift in equatorial surface zonal wind stress, compared to ERA-Interim wind stresses, which were applied directly in ORAS4 (Balmaseda *et al.*, 2013). Errors develop rapidly, and become fairly steady after around 10 days, when they can reach 25–50% of climatological wind stresses. Therefore, after around 10 days the bias correction field, which was derived for ERA-Interim wind stress forcing, is no longer valid. Its effect on the forecasts of PERS will vary regionally, depending on the structure of the correction field relative to the model drifts. Note that wind stress drifts on this time-scale are comparable to the formation of westerly wind bursts, so can be expected to impact tropical dynamics (e.g. Philander, 1981; Latif *et al.*, 1988; Fedorov and Philander, 2001), in addition to initialization shocks.

In the western and central Pacific and the Indian Ocean, wind stress drift is negative, denoting strengthening easterly winds. This raises the thermocline around the dateline, via Ekman suction, as seen in Figure 5, generating an upwelling Kelvin wave which propagates eastward. By referring to Figure 1 it can be seen that the bias correction term acts in the opposite direction, increasing the depth of the thermocline, in the central Pacific (170–180°E), and to some extent the eastern Indian Ocean (80–90°E). In these regions, the mean drift in thermocline depth in the first month is reduced in PERS through the continued application of the bias correction term (e.g. Figure 2(a)), and the Kelvin wave generated is weaker than in OP as a result. Similarly, in the central equatorial Atlantic (∼40°W), the shallowing effect of the bias correction term acts against the deepening that is caused by weakened wind stresses. In contrast, in the western and eastern edges of the Atlantic Ocean, the bias correction acts in the same direction as the initial drifts in wind stress, with respect to the thermocline, thereby accelerating the drifts.

In contrast to Niño-4, the Niño-3 (150–90°W, 5°N–5°S) thermocline initially adjusts downwards, with approximately twice the magnitude in OP as in the other forecasts, as

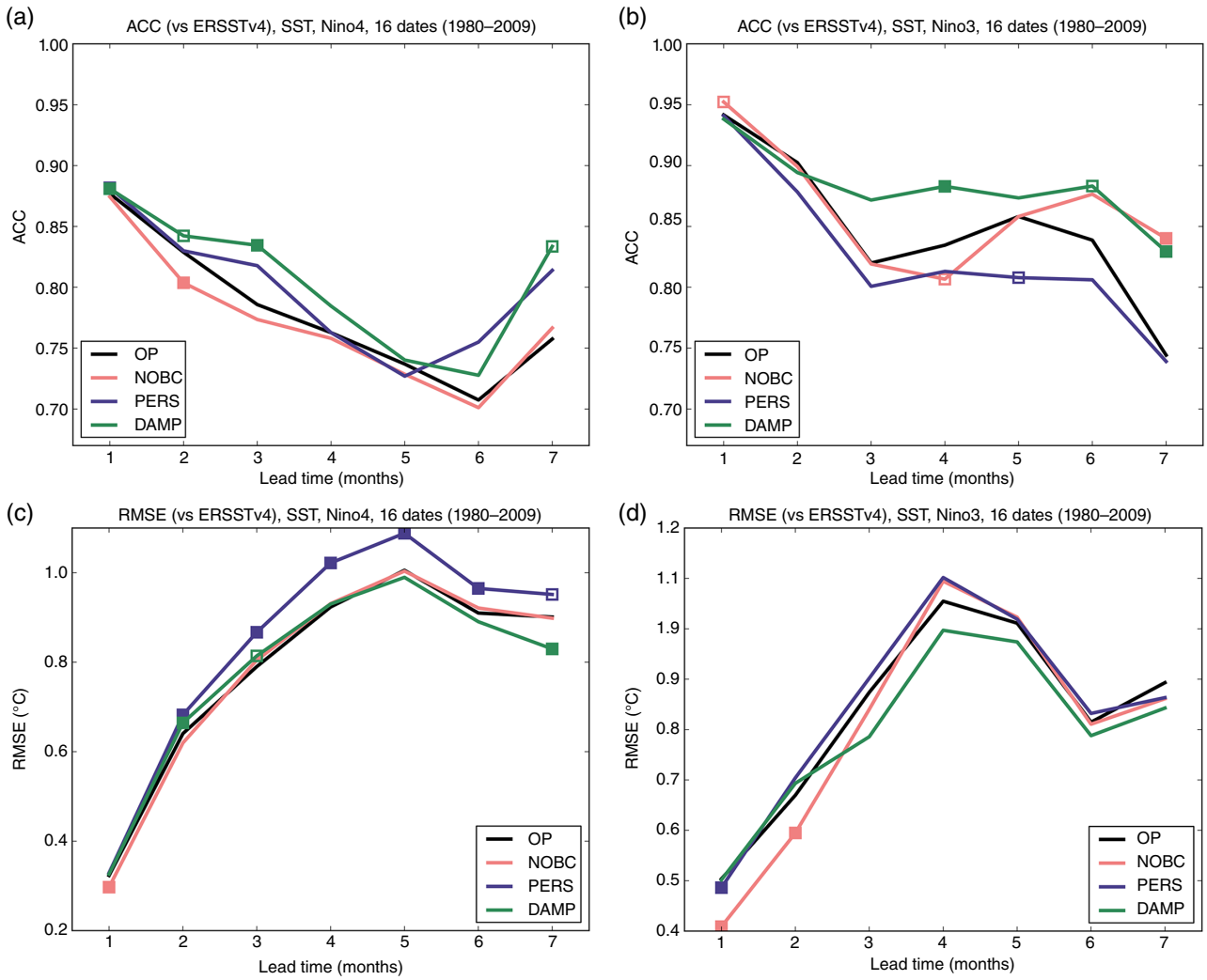

Figure 6. SST ACC versus ERSSTv4 for (a) Niño-4 and (b) Niño-3, for all four forecast sets. Filled (open) squares show values that are different from the OP value at the 95% (90%) significance level, calculated using the bootstrap method. (c, d) are as (a, b), but show raw RMSE (°C) versus ERSSTv4.

the shallowing effect of the bias correction term is removed. Continuing the bias correction reduces initial drift errors in the first few days in the Niño-3 thermocline depth in PERS and DAMP. However, the longer-term drift in Niño-3 is a shallowing of the thermocline, beginning around month 2, caused by the arrival of an upwelling Kelvin wave generated in the western Pacific, following the strengthening of easterly wind stresses there. This shallowing of 5–10 m is then amplified if the bias correction term is continually applied, resulting in a 1–2 m shallower thermocline in months 3–7 in PERS (not shown).

In summary, on the monthly time-scale, the continued application of bias correction in PERS interacts with coupled model drift and may improve or degrade the upper ocean forecast, depending on the region. In the first 10 days, the bias correction term can either amplify or dampen Kelvin wave signals generated along the thermocline by rapid drifts in surface wind stress. In the Niño-4 region, where the largest initialization shocks occur in OP, the bias correction term partially cancels the effects of wind stress drift in the short term, which may be of benefit to the forecast. This effect occurs in both PERS and DAMP, but only DAMP avoids drift interactions on longer time-scales.

3.3. *Effect on forecast skill*

We now examine whether the responses in the equatorial thermocline to initialization shocks and persisted bias correction have any longer term impact on the skill of the coupled forecasts.

The ACC for the monthly SST forecasts, evaluated against the Extended Reconstructed Sea Surface Temperature (ERSSTv4), is shown for the Niño-4 and Niño-3 regions in Figure 6(a) and (b), respectively. In Niño-3, where the most substantial area of high skill beyond ∼3 months' lead time exists, ACC for OP and NOBC are fairly similar over months 1 to 6. In fact, in the first month ACC is larger in NOBC, significant at the 90% level, and in the last 2 months ACC in NOBC again rises above that of OP, reaching 95% significance in month 7. In Niño-4, OP and NOBC are very similar from month 4 onwards. Therefore, despite the more accurate initial conditions used in OP, we suggest that the initialization shock from the removal of the bias correction has prevented significantly more skilful forecasts being achieved in the equatorial Pacific.

Over all 7 months, SST forecast skill in Niño-3 is highest in DAMP. ACC in DAMP is superior to OP from month 3 onwards, at the 95% level in months 4 and 7 (in month 7, differences are more than 0.05), and at the 90% significance level in month 6. DAMP is also superior to NOBC in months 3 to 6, with this difference being significant in months 3 and 4 (not marked). In Niño-4, DAMP again performs best, particularly in the first 3 months.

Figure 6(c) and (d) show the uncalibrated SST RMSE (i.e. including the mean drift), in Niño-4 and Niño-3, respectively. Mean error in Niño-3 is significantly lower in NOBC in the first 2 months (Figure 6(d)), perhaps due to better agreement between ORAS4_nobc and ERSSTv4 in this region than between ORAS4 and ERSSTv4, and this may contribute to the increased ACC in NOBC in month 1. However, the increased skill in NOBC in months 6 to 7 (Figure 6(b)) cannot similarly be attributed to reduced mean drift, since the mean state appears to be largely independent of initialization procedure by this stage in the

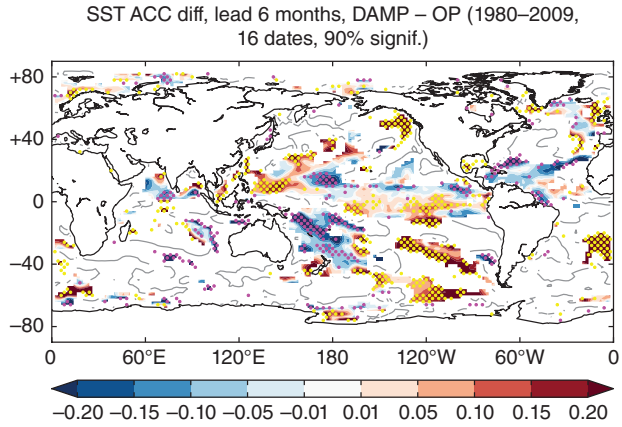

Figure 7. SST ACC difference at 6 months' lead time for DAMP minus OP. ACC differences are coloured only where the absolute ACC is at least 0.5 in one of the sets; a grey contour is drawn at values of zero outside these regions. Yellow (purple) dots mark positive (negative) differences that are significant at the 90% level, calculated using the bootstrap method.

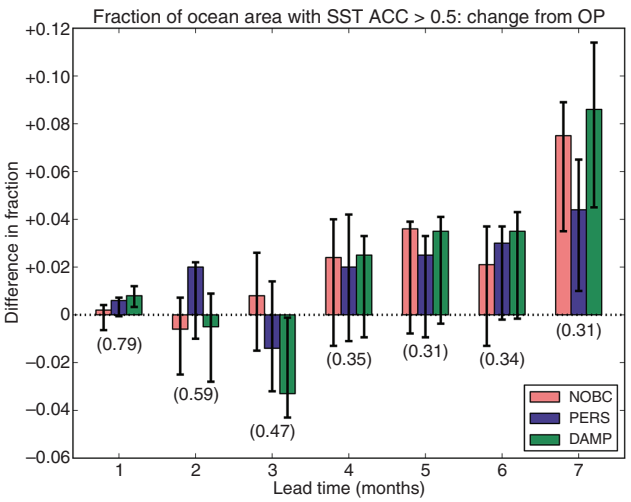

Figure 8. Difference in the fraction of ocean gridpoints with SST ACC above 0.5, between each of the modified forecast sets and OP, using ERSSTv4 as the reference. Error bars show the 90% confidence intervals for these differences, as computed using the bootstrap method. The OP fractions for each month are shown in brackets.

forecasts. Similarly, in Niño-4, the increase in skill in DAMP in months 2 and 3 (Figure 6(a)) occurs despite slightly increased mean error (Figure 6(c)). The mean drift is also decreased slightly in DAMP in Niño-3 (Figure 6(d)) from month 3 onwards, although not significantly so, which likely contributes to the improved ACC skill at these lead times.

PERS performs worst of the four methods in Niño-3 (Figure 6(b)), although differences from OP are only significant in one of the 7 months. It does somewhat better in Niño-4 (Figure 6(a)), despite a significantly larger mean drift from month 2 onwards (Figure 6(c)), the occurrence of which shows that the favourable interaction between the bias correction and the mean drift that was seen initially in the Niño-4 thermocline depth (Figure 2(a)) is not felt in SST at longer lead times.

The spatial distribution of SST ACC differences at 6 months' lead time is shown for DAMP relative to OP in Figure 7. Differences are regionally dependent, but there is an overall predominance of positive values. The regional variability is probably due in part to an insufficient number of start dates, although it is plausible for skill to be degraded by the use of bias correction in the first 20 days, in areas where this increases the mean drift in DAMP relative to OP. More generally, the reduction in initialization shock in DAMP would be expected to increase forecast skill, if anything, and this is indeed the case on average (global mean ACC is 6% higher in DAMP than in OP), and in the tropical Pacific in particular, where SST skill is most important for driving atmospheric teleconnections.

Since differences in skill at the gridpoint scale may suffer from undersampling of forecast start dates in the 16-date sets used here, it is better to compare performance on a globally averaged basis. Table 1 shows several such metrics for SST ACC in month 6. OP performs best in none of the categories, though often differences between methods are small. Overall, DAMP performs best, including showing the largest global mean skill at this lead time. PERS also performs well in several of the metrics, but does poorly in the Niño-3 region. In month 7 (not shown), the performance of DAMP relative to the other methods increases further in several metrics.

The four methods are compared globally at other lead times in Figure 8, which shows the fraction of ocean gridpoints, weighted by area, with SST ACC greater than 0.5 in each forecast month, relative to the fraction above 0.5 in OP. By this measure, NOBC develops an advantage over OP at around month 4, and maintains this advantage at subsequent lead times, although the difference only reaches the 90% confidence level in month 7. DAMP, however, emerges as at least as good as any other method from month 4 onwards, and its superiority over OP increases over months 5 to 7. OP ranks last of the four sets in each month from month 4 onwards, although by this metric it is only inferior at the 90% significance level in month 7.

The ACC was also calculated for atmospheric variables such as precipitation rate and 500 hPa geopotential averaged over months 5–7, measured against ERA-Interim, but estimates were noisy and most differences were not statistically significant. A larger number of forecasts would be needed to assess differences in skill in these fields. Since forecast skill in atmospheric variables is expected to be largely derived from skilful predictions of SST, it is still an important step to demonstrate the improvements in SST here.

Finally, the four forecast sets may be compared using the El Niño–Southern Oscillation (ENSO) forecasting 'figure of merit' (FOM), defined as the average mean absolute error in predicted SST (in °C) over months 1–6, calculated for Niño-4, Niño-3.4 (120°W–170°W, 5°S–5°N) and Niño-3, added together and multiplied by 1000 (Molteni *et al.*, 2011). To calculate this, the forecasts were first calibrated by removing the mean drift, calculated separately for each of the four seasons. The FOM for OP, NOBC, PERS and DAMP respectively are 785, 793, 834 and 739. With this metric (which extends only to month 6), it can be seen that OP and NOBC are roughly equivalent, but both are clearly outperformed by DAMP. The improvement in FOM of ∼40 points, achieved through the Niño-3 and Niño-3.4 components, is comparable to the differences seen between successive versions of the ECMWF operational system (Molteni *et al.*, 2011). These FOM values cannot be compared directly to the operational scores given by Molteni *et al.* (2011) due to the different start dates involved in each calculation, and several other differences in experimental set-up (ensemble size and generation method, use of separate calibration hindcasts).

3.4. *Understanding the improvement in DAMP*

Figure 9 shows the mean difference in 20 °C isotherm depth (averaged over 5°N–5°S) between OP and DAMP. Eastward propagating signals are present in the first month, consistent with the bias correction field in Figure 1; that is, an upwelling signal originating around 170°E, and downwelling signals originating around 50°E, 130°E, 150°W and 40°W. These signals capture the component of the thermocline shocks present in OP but not in DAMP. These signals propagate at ∼50° (month)$^{-1}$, so can be identified as Kelvin waves, and together they affect virtually all longitudes within ∼50 days. There is also evidence of slower, westward Rossby wave propagation, in the Pacific: a downwelling

Table 1. Forecast ACC for SST at 6 months' lead time in each forecast set, averaged, after applying a Fisher transformation, over the entire globe and over the Tropics (20°N–20°S), and for Niño-3 (150–90°W, 5°N–5°S) and Niño-4 (160°E–150°W, 5°N–5°S) average anomalies.

Forecast set	Global mean	Tropical mean	Niño-3	Niño-4	Fraction > 0.5	Fraction > 0.7
OP	0.38	0.50	0.84	0.71	0.34	0.14
NOBC	0.38	0.48	0.88	0.70	0.36	**0.14**
PERS	0.39	**0.51**	0.81	**0.76**	0.37	0.12
DAMP	**0.40**	0.51	**0.88**	0.73	**0.37**	0.13

ACC is calculated using ERSSTv4 as the reference for all forecast sets. The highest value in each column is in bold.
The last two columns show the fraction of ocean gridpoints (on a 2° × 2° grid) at which ACC exceeds 0.5 and 0.7.

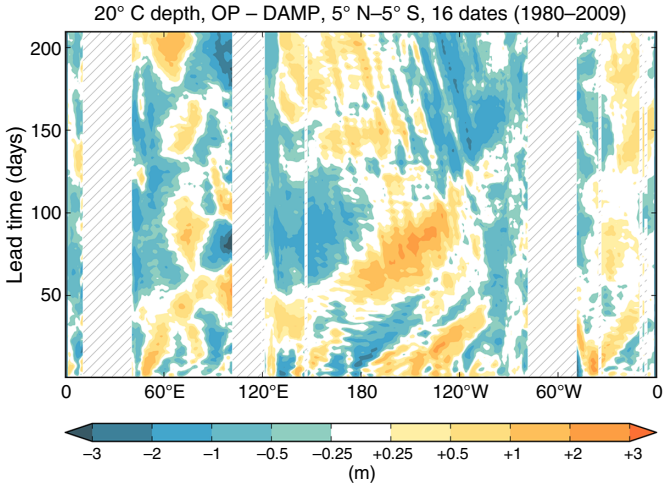

Figure 9. Equatorial (5°N–5°S) average differences in 20 °C isotherm depth (m) for OP minus DAMP (showing the additional shock present in OP), averaged over all 16 forecasts. Land areas are hatched.

signal originates at around 120°W and moves westward at ∼10° (month)$^{-1}$, passing Kelvin waves moving eastward at around 40 and 70 days' lead time. Any interactions between these waves, or between waves close to the surface and the atmosphere via convective coupling (Straub and Kiladis, 2002), will be highly nonlinear and cannot be corrected for by post-processing drift removal, and therefore have the potential to degrade forecast skill.

In Figure 10 the individual Niño-3 SST forecasts that are combined to produce the skill scores of Figure 6(b) are examined in more detail. Differences in forecast SST anomalies are small, but it can still be seen by eye that DAMP ensemble means agree slightly better with the ERSSTv4 reference than do OP ensemble means. In a few cases (February 2007, May 1984, May 1988), DAMP forecasts are more accurate throughout, possibly benefitting from a reduced initialization shock locally (the OP shock is not limited to the western Pacific: Figure 3), while in other cases (August 1980, November 1993) the DAMP advantage emerges only in the last 2 months, as a loss of accuracy in OP becomes apparent. These latter cases perhaps suggest a more complicated route for differences to reach the surface in the eastern Pacific. However, the spread of ensemble members indicates that differences in single forecasts should not be treated as robust. The correlation of SST anomalies with ERSSTv4 show particular improvements in DAMP over OP in the May forecasts, but DAMP values are higher in all four seasons. Calibrated (i.e. after mean drift removal) RMSE values also show a general improvement in DAMP, with largest differences in May. Since there are only four cases in each season, these differences should not be over-interpreted.

4. Discussion

4.1. *Impact of initialization shock*

Chen *et al.* (1997) noted that high-frequency signals in the initial conditions act as noise to the coupled model, and degrade forecast skill. They found that, while SST forecast skill at lead times of $\lesssim$8 months was improved by reduction in large-scale systematic errors in the initial conditions (as is the case in OP, compared to NOBC), improvement in skill at longer lead times was due primarily to the reduction of random noise in the initial conditions. It is in this context that the effectiveness of PERS and, in particular, DAMP can be understood. By constraining upper ocean adjustment over the first 20 days, spurious variability at 4–8 days' period is avoided in DAMP, increasing the fraction of the ocean's total spectral power that is contained in low-frequency modes, compared to OP (Figure 4). Accurate representation of these large-scale patterns can lead to improved forecast skill, particularly at long lead times.

Thoma *et al.* (2015) reported better surface temperature forecast skill at lead times of 2–9 years, especially in the Pacific, when initializing their coupled model with observed wind stress anomalies only, compared to an OP-style initialization using ORAS4. It appears that these differences can be attributed to a combination of initialization shocks due to atmosphere–ocean initial condition imbalance and bias correction removal in the OP-style initialization, which affect simulated ENSO and Interdecadal Pacific Oscillation variability on multiannual timescales. The results of Thoma *et al.* are therefore further evidence of the potential to improve long-range forecasts by reducing initialization shocks, and suggest that the benefits of DAMP may be seen more clearly still at lead times longer than 7 months.

4.2. *Significance of increases in ACC*

The non-monotonic form of the curves plotted in Figure 6(a) and (b), in comparison to the relatively smooth curves shown in Figure 5.4.5(a) of Molteni *et al.* (2011), is a result of the limited number (16) of start dates used in the ensemble. The significance test used gives an estimate of the confidence in the improvement in DAMP over the other sets, but this could be an underestimate if the 16-date ensemble is not fully representative of the period covered. However, further confidence in DAMP is provided by the fact that the DAMP ACC in Niño-3 are consistently higher than those of OP, and that this difference broadly increases with increasing lead time, consistent with the notion of errors gradually accumulating through nonlinear interactions in the upper ocean and at the surface. Also the consistent ranking of DAMP above OP, and often above NOBC, in the other metrics presented in Table 1 and Figure 8, and the clear demonstration of reduced thermocline noise in DAMP, justify the claims of improvement.

Müller *et al.* (2005) and Shi *et al.* (2015) suggested that hindcast sets of around 20 start dates are too small to give robust estimates of seasonal forecast skill. However, while individual values in Figure 6(a) and (b) may not be robust estimates of absolute skill levels, greater confidence can be given to differences between the methods, since the same forecast model is used in all cases, such that the variation in potential predictability among forecast dates should be similar for each method. DAMP performs consistently better than OP in various subsets of the 16 start dates used (significance measures in Figure 6(a) and (b), and seasonal breakdowns in Figure 10). Since differences between the OP and DAMP systems are small, differences in forecast SST on a given date arguably provide more information on likely improvements in true forecast skill than would differences between two different models on the same date.

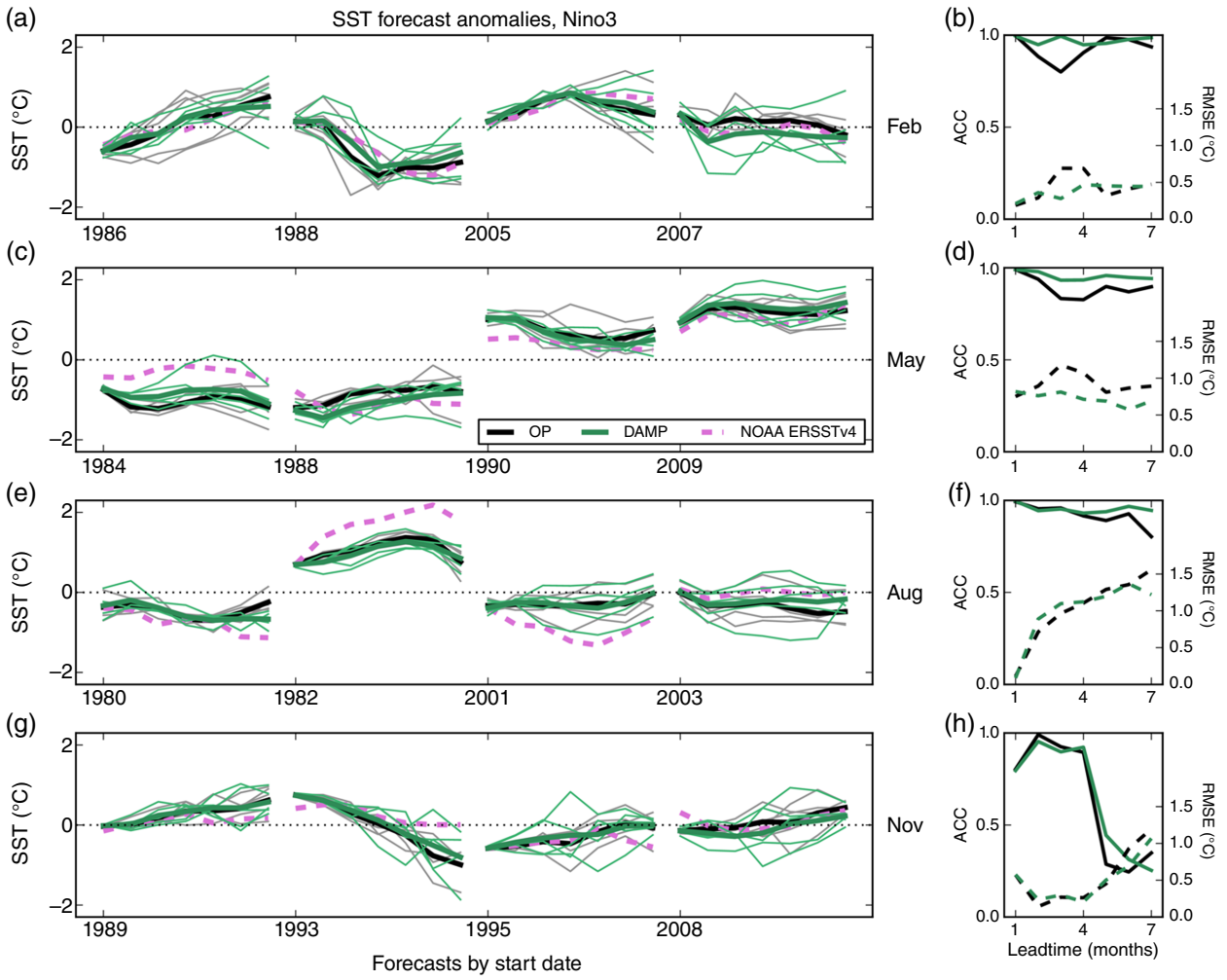

Figure 10. (a, c, e, g) Individual Niño-3 SST forecast anomalies (°C) for the 16 dates for OP and DAMP (ensemble members as thin solid lines, ensemble means as thick solid lines), and ERSSTv4 (dashed lines). (b, d, f, h) Correlation (solid lines, left axis) and calibrated RMSE (dashed lines, °C, right axis) between the forecast and ERSSTv4 anomalies, for ensemble mean forecasts grouped by season.

Nevertheless, this new method of handling bias during seasonal forecasts needs to be tested in a larger set of hindcasts to confirm its usefulness for operational prediction. More extensive tests should also investigate how differences between the methods change when larger ensembles are used (ECMWF seasonal forecasts currently use ensembles of 51 members), and how the dynamics of the initialization shock are affected by ocean model resolution.

4.3. Further improvements

Several operational centres are moving towards coupled data assimilation methods (Laloyaux *et al.*, 2016; Lea *et al.*, 2015) for producing initial conditions for seasonal and shorter-term forecasts. With a coupled analysis, it should be possible to produce a bias correction field more appropriate to the atmospheric model component of the coupled analysis. Tests using the Coupled ECMWF ReAnalysis system (CERA; Laloyaux *et al.*, 2016) have shown that tropical wind drifts are slower than in all four sets featured here, due to the use of consistent atmospheric model versions in the analysis and forecast phases, and the use of wind stresses accounting for surface ocean currents (which is not possible in an uncoupled atmospheric analysis such as ERA-Interim). With winds that drift more slowly, the bias correction should remain valid for a longer time during the forecast, so an initialization similar to DAMP or PERS could potentially be even more effective when combined with coupled assimilation.

We have not fully explored how forecast skill varies with the time-scale used to dampen the bias correction, and 20 days may not be the optimum value. For coupled initialization, the time-scale may perhaps be lengthened to further reduce the amplitude of the initialization shock. It may be desirable to vary the time-scale regionally, depending on the drifts that occur in the first month or so (Figure 5). It is also possible that a nonlinear damping of the correction field may be more suitable.

5. Summary

A number of seasonal forecast strategies, using ocean initial conditions both with and without equatorial bias correction, have been evaluated. It was found that a straightforward use of pressure correction during the ocean analysis phase followed by removal at the beginning of the forecast (set OP), as is current operational practice, leads to the generation of an initialization shock at the equatorial thermocline. It was further shown that this reduces SST forecast skill at lead times of 3 to 7 months. By some measures, using initial conditions formed without bias correction (set NOBC) outperforms this method, but better performance can be achieved using bias-corrected initial conditions by avoiding the sudden removal of the correction term. Continuing to apply the correction indefinitely (PERS) gives at least comparable forecast skill to the operational method but affects the long-term drift, while slowly removing the correction term over the first 20 days (DAMP) performs best of the four methods overall. This largely avoids the generation of noise in the thermocline and leads to more skilful tropical SST forecasts at lead times of 3 to 7 months. The results highlight the importance of the tropical ocean to delivering skilful forecasts on seasonal time-scales, and of the potential for

unwanted nonlinear interactions among propagating subsurface waves to hinder forecasting efforts. It is recommended that the method DAMP be tested over a larger set of forecast start dates to robustly measure its potential for use in operational seasonal forecasting.

Acknowledgements

The authors thank two anonymous reviewers for their constructive comments. Wavelet software used for Figure 4 was provided by C. Torrence and G. Compo, adapted for Python by E. Predybaylo, is available at http://paos.colorado.edu/research/wavelets/ (accessed 3 August 2016), and is gratefully acknow-ledged. These experiments were carried out as part of ECMWF Special Project spgbhain. Data are available on request. The work was funded by the UK Natural Environment Research Council (project ERGODICS). K. Haines is partly supported by NERC under the National Centre for Earth Observation.

Appendix

Calculation of ACC and significance testing

The centred version of the anomaly correlation coefficient (ACC) is used, in which anomalies are calculated relative to specific climatologies for each forecast or reanalysis set. Forecast and reanalysis (ORAS4 or ORAS4_nobc) ensemble means for each of the four seasons are used as seasonal climatologies, with respect to which anomalies are computed, at each lead time. Since each season of the climatologies is comprised of only 4 years (4 of the 16 dates), these will be only approximations to the true, long-term climatologies, which could only be obtained using a greater number of start dates for each forecast method. Because of this (as well as other specifics including ensemble size), the resulting ACC values cannot be compared directly to other published values calculated using larger forecast sets, such as those in Molteni *et al.* (2011).

Instead, ACC are compared among the different forecast methods. In order to account for sampling error due to the finite number of start dates used, a bootstrap method (e.g. Smith *et al.*, 2013) is used to calculate significance levels for difference between OP and the other methods at each lead time. For this, the 16-date set is sampled randomly with replacement, to form 1000 different possible forecast sets. Differences between methods are computed for each of these 1000 combinations of start dates, and differences that appear in at least 900 (950) cases are marked as being significant at the 90% (95%) level. However, this sampling method cannot account for possible unsampled climate variability in the four dates used for each season, so differences between values must still be treated with some caution.

Note also that in operational use, climatologies can in fact only be calculated using all forecasts except the one being measured ('leave-one-out'), since it has not yet been verified. By including the forecast date being measured in our calculated climatologies, ACC values will be slightly optimistic. The alternative was to calculate climatologies using only three dates for each season, so all four were used instead to increase the stability of the climatologies. Since ACC values are only being compared among the forecast methods in this work, overestimation of their absolute values (in all cases) is not a problem.

References

Balmaseda M, Anderson D. 2009. Impact of initialization strategies and observations on seasonal forecast skill. *Geophys. Res. Lett.* **36**: L01701, doi: 10.1029/2008GL035561.

Balmaseda MA, Dee DP, Vidard A, Anderson DLT. 2007. A multivariate treatment of bias for sequential data assimilation: Application to the tropical oceans. *Q. J. R. Meteorol. Soc.* **133**: 167–179.

Balmaseda MA, Mogensen K, Weaver AT. 2013. Evaluation of the ECMWF ocean reanalysis system ORAS4. *Q. J. R. Meteorol. Soc.* **139**: 1132–1161.

Barnston AG, Li S, Mason SJ, DeWitt DG, Goddard L, Gong X. 2010. Verification of the first 11 years of IRI's seasonal climate forecasts. *J. Appl. Meteorol. Climatol.* **49**: 493–520.

Barsugli JJ, Sardeshmukh PD. 2002. Global atmospheric sensitivity to tropical SST anomalies throughout the Indo-Pacific basin. *J. Clim.* **15**: 3427–3442.

Bell MJ, Martin MJ, Nichols NK. 2004. Assimilation of data into an ocean model with systematic errors near the Equator. *Q. J. R. Meteorol. Soc.* **130**: 873–893.

Blockley EW, Martin MJ, McLaren AJ, Ryan AG, Waters J, Lea DJ, Mirouze I, Peterson KA, Sellar A, Storkey D. 2014. Recent development of the Met Office operational ocean forecasting system: An overview and assessment of the new Global FOAM forecasts. *Geosci. Model Dev.* **7**: 2613–2638, doi: 10.5194/gmd-7-2613-2014.

Chen D, Zebiak SE, Cane MA, Busalacchi AJ. 1997. Initialization and predictability of a coupled ENSO forecast model. *Mon. Weather Rev.* **125**: 773–788.

Dee DP, Uppala SM, Simmons AJ, Berrisford P, Poli P, Kobayashi S, Andrae U, Balmaseda MA, Balsamo G, Bauer P Bechtold P, Beljaars ACM, van de Berg L, Bidlot J, Bormann N, Delsol C, Dragani R, Fuentes M, Geer AJ, Haimberger L, Healy SB, Hersbach H, Hólm EV, Isaksen L, Kållberg P, Köhler M, Matricardi M, McNally AP, Monge-Sanz BM, Morcrette J-J, Park B-K, Peubey C, de Rosnay P, Tavolato C, Thépaut J-N, Vitart F. 2011. The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.* **137**: 553–597.

Fedorov AV. 2002. The response of the coupled tropical ocean–atmosphere to westerly wind bursts. *Q. J. R. Meteorol. Soc.* **128**: 1–23.

Fedorov AV, Philander SG. 2001. A stability analysis of tropical ocean-atmosphere interactions: Bridging measurements and theory for El Niño. *J. Clim.* **14**: 3086–3101.

Harrison D, Giese BS. 1988. Remote westerly wind forcing of the eastern equatorial Pacific; some model results. *Geophys. Res. Lett.* **15**: 804–807, doi: 10.1029/GL015i008p00804.

Hayes S, Mangum L, Picaut J, Sumi A, Takeuchi K. 1991. TOGA-TAO: A moored array for real-time measurements in the tropical Pacific Ocean. *Bull. Am. Meteorol. Soc* **72**: 339–347.

Huang B, Banzon VF, Freeman E, Lawrimore J, Liu W, Peterson TC, Smith TM, Thorne PW, Woodruff SD, Zhang H-M. 2015. Extended reconstructed sea surface temperature version 4 (ERSST v4). Part I: Upgrades and intercomparisons. *J. Clim.* **28**: 911–930.

Kent EC, Fangohr S, Berry DI. 2013. A comparative assessment of monthly mean wind speed products over the global ocean. *Int. J. Climatol.* **33**: 2520–2541.

Laloyaux P, Balmaseda M, Dee D, Mogensen K, Janssen P. 2016. A coupled data assimilation system for climate reanalysis. *Q. J. R. Meteorol. Soc.* **142**: 65–78, doi: 10.1002/qj.2629.

Latif M, Biercamp J, Von Storch H. 1988. The response of a coupled ocean-atmosphere general circulation model to wind bursts. *J. Atmos. Sci.* **45**: 964–979.

Lea D, Mirouze I, Martin M, King R, Hines A, Walters D, Thurlow M. 2015. Assessing a new coupled data assimilation system based on the Met Office coupled atmosphere–land–ocean–sea ice model. *Mon. Weather Rev.* **143**: 4678–4694.

MacLachlan C, Arribas A, Peterson KA, Maidens A, Fereday D, Scaife AA, Gordon M, Vellinga M, Williams A, Comer RE, Camp J, Xavier P, Madec G. 2014. Global Seasonal Forecast System version 5 (GloSea5): A high-resolution seasonal forecast system. *Q. J. R. Meteorol. Soc.* **141**: 1072–1084, doi: 10.1002/qj.2396.

Magnusson L, Alonso-Balmaseda M, Corti S, Molteni F, Stockdale T. 2013. Evaluation of forecast strategies for seasonal and decadal forecasts in presence of systematic model errors. *Clim. Dyn.* **41**: 2393–2409.

Mason SJ, Goddard L. 2001. Probabilistic precipitation anomalies associated with ENSO. *Bull. Am. Meteorol. Soc.* **82**: 619–638.

Mathieu P, Sutton R, Dong B, Collins M. 2004. Predictability of winter climate over the North Atlantic European region during ENSO events. *J. Clim.* **17**: 1953–1974.

Meehl GA, Goddard L, Murphy J, Stouffer RJ, Boer G, Danabasoglu G, Dixon K, Giorgetta MA, Greene AM, Hawkins E Hegerl G, Karoly D, Keenlyside N, Kimoto M, Kirtman B, Navarra A, Pulwarty R, Smith D, Stammer D, Stockdale T. 2009. Decadal prediction: Can it be skilful? *Bull. Am. Meteorol. Soc.* **90**: 1467–1485.

Mogensen KS, Balmaseda MA, Weaver A. 2012. *The NEMOVAR Ocean Data Assimilation System as Implemented in the ECMWF Ocean Analysis for System 4*, Technical Memorandum 668. ECMWF: Reading, UK.

Molteni F, Stockdale T, Balmaseda M, Balsamo G, Buizza R, Ferranti L, Magnusson L, Mogensen K, Palmer TN, Vitart F. 2011. *The New ECMWF Seasonal Forecast System (System 4)*, Technical Memorandum 656. ECMWF: Reading, UK.

Mulholland DP, Laloyaux P, Haines K, Balmaseda MA. 2015. Origin and impact of initialization shocks in coupled atmosphere–ocean forecasts. *Mon. Weather Rev.* **143**: 4631–4644.

Müller W, Appenzeller C, Schär C. 2005. Probabilistic seasonal prediction of the winter North Atlantic Oscillation and its impact on near-surface temperature. *Clim. Dyn.* **24**: 213–226.

Philander S. 1981. The response of equatorial oceans to a relaxation of the trade winds. *J. Phys. Oceanogr.* **11**: 176–189.

Pierce DW, Barnett TP, Tokmakian R, Semtner A, Maltrud M, Lysne J, Craig A. 2004. The ACPI Project, Element 1: Initializing a coupled climate model from observed conditions. *Clim. Change* **62**: 13–28, doi: 10.1023/B:CLIM.0000013676.42672.23.

Polkova I, Köhl A, Stammer D. 2014. Impact of initialization procedures on the predictive skill of a coupled ocean–atmosphere model. *Clim. Dyn.* **42**: 3151–3169, doi: 10.1007/s00382-013-1969-4.

Rahmstorf S. 1995. Climate drift in an ocean model coupled to a simple, perfectly matched atmosphere. *Clim. Dyn.* **11**: 447–458.

Richardson LF. 1922. *Weather Prediction by Numerical Process*. Cambridge University Press: Cambridge, UK. Reprinted 1965 by Dover: New York, NY.

Saha S, Moorthi S, Wu X, Wang J, Nadiga S, Tripp P, Behringer D, Hou Y-T, Chuang H-Y, Iredell M Ek M, Meng J, Yang R, Peña Mendez M, van den Dool H, Zhang Q, Wang W, Chen M, Becker E. 2014. The NCEP climate forecast system version 2. *J. Clim.* **27**: 2185–2208.

Sanchez-Gomez E, Cassou C, Ruprich-Robert Y, Fernandez E, Terray L. 2016. Drift dynamics in a coupled model initialized for decadal forecasts. *Clim. Dyn.* **46**: 1819–1840.

Shi W, Schaller N, MacLeod D, Palmer T, Weisheimer A. 2015. Impact of hindcast length on estimates of seasonal climate predictability. *Geophys. Res. Lett.* **42**: 1554–1559, doi: 10.1002/2014GL062829.

Smith DM, Cusack S, Colman AW, Folland CK, Harris GR, Murphy JM. 2007. Improved surface temperature prediction for the coming decade from a global climate model. *Science* **317**: 796–799.

Smith DM, Eade R, Pohlmann H. 2013. A comparison of full-field and anomaly initialization for seasonal to decadal climate prediction. *Clim. Dyn.* **41**: 3325–3338.

Smith PJ, Fowler AM, Lawless AS. 2015. Exploring strategies for coupled 4D-Var data assimilation using an idealised atmosphere–ocean model. *Tellus A* **67**: 27025, doi: 10.3402/tellusa.v67.27025.

Stockdale TN. 1997. Coupled ocean–atmosphere forecasts in the presence of climate drift. *Mon. Weather Rev.* **125**: 809–818.

Straub KH, Kiladis GN. 2002. Observations of a convectively coupled Kelvin wave in the eastern Pacific ITCZ. *J. Atmos. Sci.* **59**: 30–53.

Thoma M, Greatbatch RJ, Kadow C, Gerdes R. 2015. Decadal hindcasts initialized using observed surface wind stress: Evaluation and prediction out to 2024. *Geophys. Res. Lett.* **42**: 6454–6461, doi: 10.1002/2015GL064833.

Torrence C, Compo GP. 1998. A practical guide to wavelet analysis. *Bull. Am. Meteorol. Soc.* **79**: 61–78.

Trenberth K, Branstator G, Karoly D, Kumar A, Lau N, Ropelewski C. 1998. Progress during TOGA in understanding and modeling global teleconnections associated with tropical sea surface temperatures. *J. Geophys. Res.* **103**: 14291–14324, doi: 10.1029/97JC01444.

Vecchi GA, Harrison D. 2000. Tropical Pacific sea surface temperature anomalies, El Niño, and equatorial westerly wind events. *J. Clim.* **13**: 1814–1830.

Weisheimer A, Doblas-Reyes F, Palmer T, Alessandri A, Arribas A, Déqué M, Keenlyside N, MacVean M, Navarra A, Rogel P. 2009. ENSEMBLES: A new multi-model ensemble for seasonal-to-annual predictions –Skill and progress beyond DEMETER in forecasting tropical Pacific SSTs. *Geophys. Res. Lett.* **36**: L21711, doi: 10.1029/2009GL040896.

Zhang S. 2011. A study of impacts of coupled model initial shocks and state-parameter optimization on climate predictions using a simple pycnocline prediction model. *J. Clim.* **24**: 6210–6226.

Zhang S, Harrison MJ, Rosati A, Wittenberg A. 2007. System design and evaluation of coupled ensemble data assimilation for global oceanic climate studies. *Mon. Weather Rev.* **135**: 3541–3564.