Evaluating Benefits of Two-Way Ocean–Atmosphere Coupling for Global NWP Forecasts

MICHAEL VELLINGA, DAN COPSEY, TIM GRAHAM, SEAN MILTON, AND TIM JOHNS

Met Office, Exeter, United Kingdom

(Manuscript received 2 March 2020, in final form 4 August 2020)

ABSTRACT: We evaluate the impact of adding two-way coupling between atmosphere and ocean to the Met Office deterministic global forecast model. As part of preoperational testing of this coupled NWP configuration we have three years of daily forecasts, run in parallel to the uncoupled operational forecasts. Skill in the middle and upper troposphere out to $T + 168$ h is generally increased compared to the uncoupled model. Improvements are strongest in the tropics and largely neutral in midlatitudes. We attribute the additional skill in the atmosphere to the ability of the coupled model to predict sea surface temperature (SST) variability in the (sub)tropics with greater skill than persisted SSTs as used in uncoupled forecasts. In the midlatitude, ocean skill for SST is currently marginally worse than persistence, possibly explaining why there is no additional skill for the atmosphere in midlatitudes. Sea ice is predicted more skillfully than persistence out to day 7 but the impact of this on skill in the atmosphere is difficult to verify. Two-way air–sea coupling benefits tropical cyclone forecasts by reducing median track and central pressure errors by around 5%, predominantly from $T + 90$ to $T + 132$ h. Benefits from coupling are largest for large cyclones, and for smaller storms coupling can be detrimental. In this study skill in forecasts of the Madden–Julian oscillation does not change with two-way air–sea coupling out to $T + 168$ h.

SIGNIFICANCE STATEMENT: In many operational weather forecasts, interactions between atmosphere and ocean are simplified to acting in one direction only: the ocean can affect the state of the atmosphere but the ocean itself is insensitive to changes in the atmosphere. We investigate the impact on forecasts out to day 7 when coupled interactions are allowed in the Met Office global forecast model. Benefits of coupling are greatest in the tropics and give a 5% reduction in track error of tropical cyclones during days 4–7. In midlatitudes, the effect of coupling is neutral, probably because the resolution of the ocean model we use is insufficient. Our study shows that interactive air–sea coupling improves weather forecasts and suggests further ways to improve coupled forecasts.

KEYWORDS: Forecast verification/skill; Numerical weather prediction/forecasting; Short-range prediction; Coupled models; Model comparison

1. Introduction

Operational numerical weather prediction (NWP) centers continually strive to advance forecast accuracy in their prediction systems. Such has been the progress over recent decades that useful predictability has been extended in NWP models by roughly one day per decade for verified atmospheric variables (Simmons and Hollingsworth 2002; Benjamin et al. 2018). One possible means to further increase skill in short-range (taken here as 1–3 days) and medium-range (3–10 days) forecasts would be to improve the forecast accuracy of sea surface temperatures (SSTs) that interact with the atmospheric models by roughly one day per decade for verified atmospheric variables (Simmons and Hollingsworth 2002; Benjamin et al. 2018). One possible means to further increase skill in short-range forecasts is to improve the representation of air–sea coupled feedbacks. The increased skill largely derives from the ability to predict tropical SST and subsequent response by the atmosphere at seasonal time scales such as PNA and NAO patterns (e.g., Wallace et al. 1990). Evidence that coupled air–sea interaction can improve key physical processes and variability in models on intraseasonal time scales (14–60 days) is now also becoming more compelling, particularly in the tropics. Forecasts of the Madden–Julian oscillation (MJO; Madden and Julian 1971)—the principal mode of tropical variability on intraseasonal time scales—might therefore be improved through an accurate representation of air–sea coupled feedbacks. The representation of the MJO in models shows sensitivity to the presence of atmosphere–ocean feedbacks (Waliser et al. 1999; Matthews 2004; Klingaman et al. 2008) and it has been shown that ocean–atmosphere coupling acts to improve the simulated spatiotemporal evolution of the MJO and phase relationship between convection and SST over the equatorial Indian Ocean and western Pacific (Kim et al. 2010; Shelly et al. 2014). Notwithstanding this growing evidence of the importance of air–sea coupling for the MJO, implications for MJO predictions up to the medium range are less obvious. Woolnough et al. (2007) found that in hindcasts for the period 15 December 1992–31 January 1993 the benefit for MJO forecasts of predicting SST, compared to persisting SST, only appears from day 10. This implies skill in days 1–9 derives mostly from initial conditions in the atmosphere with a more limited role for variations in SST and associated air–sea interaction processes.

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/WAF-D-20-0035.s1.

Corresponding author: Michael Vellinga, michael.vellinga@metoffice.gov.uk

DOI: 10.1175/WAF-D-20-0035.1

For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).
Furthermore, Fu et al. (2015) show in their analysis of the Dynamics of the Madden–Julian Oscillation (DYNAMO) field campaign that the importance of air–sea interaction varies substantially between MJO events. The question remains, therefore, how much we can expect MJO predictions to benefit on average from interactive air–sea coupling and at what lead times. Addressing this will be one of the aims of this paper.

Tropical cyclones (TCs), whose high wind speeds make them efficient mixers of the upper ocean (Jacob et al. 2000), drawing cooler subthermoline waters upward and thereby cooling the surface, exemplify strong air–sea interaction. Upper oceanic mixed layer heat content may be a key determinant in predicting hurricane intensity (Goni and Trinanes 2003; Scharroo et al. 2005; Shay et al. 2000). Although TC predictions in global NWP systems have seen a gradual improvement in track forecast errors over the last 25 years (Yamaguchi et al. 2017), they generally suffer from underpredicting intensity. There are many possible causes for this, but model studies have shown sensitivity to resolution and representation of convection (Short and Petch 2018) and initialization (Heming 2016). This sensitivity to model resolution could explain why more recently model upgrades (involving increased horizontal resolution) have caused some global models to now overestimate minima of TC central pressure minima in some cases (Heming 2016; Mogensen et al. 2017). In the latter study it was found for one TC prediction case that including interactive air–sea coupling resulted in a significant reduction of overdeepening, compared to the uncoupled prediction. It is important to examine this point in a larger number of TC cases. This is another aim of our paper.

Observational studies suggest that air–sea interaction in the midlatitudes is scale dependent (Chelton et al. 2004, 2001; Park et al. 2006). Over large spatial areas and longer time scales, atmospheric forcing of oceanic variability tends to dominate at the air–sea interface (Liu et al. 1994). However, on the mesoscale, in the region of shorter-lived features such as ocean eddies and SST fronts, the ocean is thought to force the atmospheric boundary layer, as evident in the positive correlation between SST and surface wind speed, a relationship that is generally negative at larger spatial scales (Chelton et al. 2004). Sharp SST gradients along oceanic fronts (such as the Gulf Stream or Kuroshio) also provide pathways for forcing of the midlatitude atmosphere, potentially well into the midtroposphere (Minobe et al. 2008; Sheldon et al. 2017). Most ocean mesoscale features have lifetimes of weeks to months (Small et al. 2008). This then raises the question of the need for predicting their variability at the much shorter time scales of short- to medium-range global NWP. To what extent this type of midlatitude air–sea interactions needs to be considered a coupled process at NWP time scales (i.e., subsequent evolution of SST during the prediction window is large enough to have a noticeable impact on the atmosphere) is currently an open question. In the tropics the arguments for this appear stronger. Apart from TCs mentioned earlier, tropical instability waves (TIWs) modify SST in the tropical Pacific and Atlantic. These waves have periods of 17 and 33 days (Lyman et al. 2007), giving them relevance in the short- to medium-range forecast window.

In operational short-range NWP forecasts SST and sea ice have traditionally been treated as prescribed, fixed lower boundary conditions to the atmosphere. In that setup, SST observations from the latest analysis available at the beginning of the forecast are persisted for the duration of the forecast. This uncoupled forecast system is referred to as UNCP LD in this study. A coupled numerical weather prediction (hereafter CPLD NWP) model that predicts SSTs more accurately than persisted SST anomalies therefore has the potential to advance forecast skill relative to conventional uncoupled NWP through a more physically consistent representation of air–sea interactions and fast atmospheric physical processes (e.g., convection) coupled to SST and sea ice distributions. Coupled global NWP at short to extended range is a relatively new development, but some centers have employed this in operational forecasts for some years (Mogensen et al. 2017; Smith et al. 2018). Results show encouraging improvements in predictions of tropical cyclone intensity (Mogensen et al. 2017; Smith et al. 2018) and reduction of large-scale errors of geopotential height (Smith et al. 2018).

Understanding the advantages of coupled over uncoupled predictions in global NWP is in its infancy. Predicting SST in coupled models can suffer from systematic errors (e.g., drift during the forecast). Coupled forecasts bring the additional task of initializing ocean and sea ice models, which has its own uncertainties. Much remains to be understood about the actual benefit for atmosphere forecast skill of coupled forecasts in the current generation of prediction systems, given these uncertainties. A key question is whether such a system that predicts SST outperforms one that uses persisted SST, in terms of its impact on atmosphere forecast skill up to the medium range.

The purpose of our paper is to contribute to this developing understanding. The Met Office is preparing implementation of an operational coupled NWP forecasting system by 2022. As part of this preparation it has been running a near-real-time coupled global NWP forecast daily since 2016, in parallel to its operational global uncoupled system. This growing set of parallel coupled and uncoupled forecasts already spans several years and provides a valuable opportunity to address some of the questions raised in the preceding literature synthesis. Specifically, by assessing performance of two-way coupled NWP relative to uncoupled NWP we will explore the following questions within the Met Office forecasting systems: 1) What is the impact of predicting SST and sea ice on atmosphere forecast accuracy out to 168 h? 2) Does the quality of forecast SST and sea ice outperform that of persistence? Other questions such as benefits of coupled NWP for ocean forecasting (Lewis et al. 2019) or challenges and benefits from coupled initialization (Browne et al. 2018; Guiavarch et al. 2019; Skachko et al. 2019) are not considered in this study. In section 2 we describe model configurations and forecast methodology. In section 3 we present a general assessment of forecast performance, concentrating on standard atmospheric verification and quality of SST and sea ice predictions. Specific aspects of forecast performance referred to above (TCs, MJO, and midlatitude circulation) are presented in section 4. A discussion follows in section 5.

2. Model configurations and forecasts

As part of the preparations for a transition of a global coupled NWP system into operations mentioned in section 1, the
Met Office has been continuously running daily forecasts with a coupled forecasting model since May 2016. The starting point for the coupled NWP model development was a coupled climate model—the HadGEM3 model (Hewitt et al. 2011), adapted to run in initialized forecast mode. The atmosphere–ocean–ice–land forecast model used here consist of the global atmosphere (version GA6.1) and global land (versions GL6.1 and GL8.1) components (Walters et al. 2017, 2019), coupled to the Nucleus for European Modelling of the Ocean (NEMO) consortium (Madec et al. 1998) ocean model versions GO5 (Megann et al. 2014) and GO6 (Storkey et al. 2018), which are based on NEMO versions 3.4 and 3.6, respectively; and the GSI6 (Rae et al. 2015) and GSI8 (Ridley et al. 2018) implementations of the Los Alamos CICE sea ice model (Hunke and Lipscomb 2010). Scientific updates were applied to the operational atmosphere and ocean forecast models approximately every 12 months. These were implemented in the coupled forecast model at the same time, to maintain traceability. Table 1 lists the model configurations used at various times.

N768 corresponds to a horizontal grid spacing of 0.234° in longitude by 0.156° in latitude. N1280 corresponds to a grid spacing of 0.14° in longitude by 0.09° in latitude (about 11 km × 10 km at 45° latitude, 15 km × 10 km at the equator). The L85 identifier refers to a vertical resolution of 85 levels, with 50 and 35 levels in the troposphere and stratosphere, respectively. It has a 20-m bottom layer and 16 levels in the lowest 2 km. L70 refers to 70 vertical levels; the main difference with L85 is reduced resolution in the stratosphere (20 instead of 35 levels). The 50 levels of L70 that lie in the troposphere are virtually identical to those at L85. The operational model uses L70 vertical resolution throughout the whole period used in this study, which is one of the few differences with the atmosphere in the coupled forecast model, until 26 September 2018. From that date L70 is also used in the CPLDNWP configuration. Another difference is the mask used by the atmosphere model to distinguish between land and sea points. In the CPLDNWP configuration the mask is defined on the 1/4° ocean grid. Coastal tiling is used in the atmosphere grid where atmosphere grid boxes overlay some land and some ocean. In the UNCPPLD model the mask is defined on the higher-resolution atmosphere grid. The atmosphere model time step is 4 (7.5) min at N1280 (N768) resolution.

The JULES land surface model (Best et al. 2011) contains four soil layers with the top layer centered at 5 cm and additional layers of increasing thickness down to 2-m depth. The NEMO ocean model component is run at 1/4° quasi-isotropic resolution on the ORCA025 tripolar horizontal grid (Bernard et al. 2006) with 75 levels in the vertical varying in thickness from 1 m at the surface to about 11, 60, and 200 m at depths of 100, 600, and 5000 m, respectively. CICE shares the ocean model’s ORCA025 horizontal grid, but with a different B-grid arrangement of prognostic variables. Both the ocean and CICE model components use a 20-min time step, with subtime stepping for the ice dynamics. In coupled simulations the atmosphere and land components exchange information with the ocean and sea ice components via the OASIS3 coupler (Valcke 2013) using an hourly coupling frequency.

The coupled model is initialized daily at 0000 UTC and run out to 10 days (to 15 days until 1 March 2018). Initial states are provided separately for initialization of atmosphere-land and ocean–sea ice components, from uncoupled data assimilation (DA) systems. The atmosphere–land initial conditions come from the Met Office 4D-Var DA system described by Rawlins et al. (2007). (referred to as “UM analyses”). UM analyses use SST and sea ice concentration analyses from OSTIA (Domon et al. 2012) for surface boundary conditions. The data-assimilation system used by OSTIA was updated in 2018 and 2019 (Fiedler et al. 2019; Good et al. 2020). This resulted in better verification against independent ocean drifter observations and improved representation of small-scale features in the SST analyses. These improvements to OSTIA are beneficial to atmosphere forecasts that use OSTIA as boundary forcing (Fiedler et al. 2019). Systematic differences between OSTIA and other SST analyses and independent drifter data are available from the NOAA SST quality monitor (SQUAM): https://www.star.nesdis.noaa.gov/sod/sst/squam/analysis/l4

The ocean and sea ice initial conditions were generated using the FOAM-NEMOVAR DA system described by Waters et al. (2014) and Blockley et al. (2014) (“FOAM analyses”). This “uncoupled” initialization approach has the advantage that it allows testing of CPLDNWP forecasts at present. More sophisticated coupled initialization methods are under development (e.g., weakly coupled DA), in which ocean and atmosphere data assimilation use a common background state from a coupled model integration. One of the advantages of weakly coupled DA is that it will reduce some of the inconsistencies we have encountered (e.g., initialization of the atmosphere over the ocean and sea ice: section 3b).

UNCPPLD forecasts are forced by analyzed SST and sea ice concentration fields from OSTIA that are kept constant. This removes all SST variability during the UNCPPLD forecast, for example, the diurnal and annual cycles.

### 3. Forecast performance

First we evaluate the general level of skill in the current CPLDNWP predictions and compare it to skill of uncoupled predictions. We quantify skill in terms of the difference of CPLDNWP predictions and compare it to skill of uncoupled predictions. We quantify skill in terms of the difference of forecast error compared to that of a reference forecast (e.g., uncoupled, or persistence forecasts).

---

<table>
<thead>
<tr>
<th>Period</th>
<th>Atmosphere resolution (horizontal; vertical)</th>
<th>Atmosphere configuration</th>
<th>Ocean configuration</th>
<th>Sea ice configuration</th>
<th>Land surface configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 May 2016–11 Jul 2017</td>
<td>N768; L85</td>
<td>GA6.1</td>
<td>GO5</td>
<td>GSI6</td>
<td>GL6.1</td>
</tr>
<tr>
<td>12 Jul 2017–25 Sep 2018</td>
<td>N1280; L85</td>
<td>GA6.1</td>
<td>GO5</td>
<td>GSI6</td>
<td>GL6.1</td>
</tr>
<tr>
<td>26 Sep 2018–25 Sep 2019</td>
<td>N1280; L70</td>
<td>GA6.1</td>
<td>GO6</td>
<td>GSI8</td>
<td>GL8.1</td>
</tr>
</tbody>
</table>
For a series of $n$ forecasts $f_i$ and corresponding verifying observations $o_i$, the mean error (ME), or bias, is defined by

$$ ME = \frac{1}{n} \sum_i (f_i - o_i), $$

mean-squared error is defined by

$$ MSE = \frac{1}{n} \sum_i (f_i - o_i)^2, $$

and root-mean-squared error (RMSE) is defined by

$$ \text{RMSE} = \sqrt{\text{MSE}}. $$

We also use mean absolute error (MAE) to quantify differences in bias between two forecasts or for spatial averages of bias:

$$ \text{MAE} = \frac{1}{n} \sum_i |(f_i - o_i)|. $$

A useful error metric for model forecast, that does not penalize for mean model bias is variance (Var) of the difference between forecast and observations. Standard decomposition (e.g., Murphy and Epstein 1989) relates this quantity to MSE and ME:

$$ \text{Var}(f - o) = \text{MSE} - \text{ME}^2. $$

Finally, for the purpose of evaluating SST and sea ice forecasts it is useful to define a skill measure $S$ as the difference between $\text{Var}(f - o)$ from the CPLDNWP system and that from a persistence forecast:

$$ \text{Skill} = S = \text{Var}(f - o)_{\text{CPLDNWP}} - \text{Var}(f - o)_{\text{persistence}}. $$

Negative $S$ indicates greater skill in the coupled forecast compared to persistence. In section 3a we use RMSE of standard verification of atmospheric variables over defined regions as a function of lead time, against UM analyses for both predictions. Note these analyses are native to UNCPLD, but not to CPLDNWP, potentially putting the latter at a disadvantage (e.g., if model biases of UNCPLD and CPLDNWP are different). We verify at 6-hourly intervals up to a lead time of 168 h, dictated by the length of the operational UNCPLD predictions. We calculate RMSE for the period DJF 2018/19. The difference in RMSE (coupled minus uncoupled) is shown in Fig. 1. We are currently unable to verify against observations in a consistent way in the coupled and uncoupled forecasts, although this capability has now been implemented and is possible for future forecasts.

Generally, NWP RMSE scores in CPLDNWP improve compared to that of the uncoupled system. This shows there is a benefit of interactive air–sea coupling out to $T + 168$, across key forecast metrics. The only field with degraded performance
is 2-m temperature (T2m). This is explored in more detail in the next section.

**b. T2m forecast skill**

At short lead times poorer verification in CPLDNWP than in UNCPLD with respect to UM analyses occurs mostly over the ocean, while over land the two forecasting systems are performing very similarly (Fig. 2). This degraded performance over the ocean at short lead times reflects the different SSTs from which the two systems start their forecast: FOAM in CPLDNWP and OSTIA in the UM. This penalizes verification against OSTIA-driven atmosphere analysis of CPLDNWP, particularly at short lead time when SST has not evolved much. The similarity between \( T_{24} \), T2m RMSE (Fig. 2a) and the RMS difference between FOAM and OSTIA (Fig. S1 in the online supplemental material) supports this. At longer lead times the SST field evolves from its initial state and CPLDNWP is able to predict this evolution with some skill (see section 3c). At \( T + 168 \) this results in large regions over the ocean with reduced T2m RMSE compared to UNCPLD. These include coastal upwelling regions of the subtropics or semienclosed basins like the Mediterranean and Black Sea. However, over some shallow basins (such as the English Channel or the northern Caspian Sea) this initial degradation of T2m skill persists throughout the forecast.

Degradation of T2m forecasts at short lead times is particularly large over sea ice, Fig. 2a. Again, as with SST, there is a discrepancy between UM analyses that use sea ice information (from OSTIA) and initialization of sea ice in CPLDNWP from FOAM analyses. This penalizes CPLDNWP for differences in its T2m predictions that arise from systematic sea ice differences between FOAM and OSTIA. Compared to SST challenges for specifying sea ice in the analyses are even greater in that sea ice fraction and sea ice thickness are required. The NEMO-CICE model in FOAM uses multicategory sea ice thickness, which is able to account for spatial inhomogeneities. However, in the UM a simple assumption of uniform ice thickness is made. These differences in treatment of sea ice lead to substantial differences (\( >100 \text{ W m}^{-2} \)) between surface heat fluxes at the sea ice–atmosphere interface in the two systems. These, in turn, cause substantial differences in T2m over sea ice. Estimates of the analyzed state of the atmosphere over sea ice improve with better representation of sea ice during the assimilation (e.g., by using coupled assimilation methods) (Guiavarch' et al. 2019; Skachko et al. 2019). Until then the apparent deterioration in T2m of CPLDNWP compared to UNCPLD (as in Fig. 2) should not be seen as evidence of poor model skill of sea ice predictions, as will be shown in section 3d.

Finally, apparent degradation of forecast skill occurs over and in the vicinity of some large lakes, like the Karelian Lakes. Temperatures of larger lakes are included in OSTIA (Fiedler et al. 2014) and are used in UNCPLD for its forecasts and analyses. This includes the Karelian Lakes. In CPLDNWP the largest lakes are part of the ocean model that predicts lake temperatures (e.g., Lake Victoria). Most lakes, however, are not resolved by NEMO at the resolution used. Because OSTIA is not used in CPLDNWP such smaller lakes are currently
treated as wet soil by the atmosphere model in CPLDNWP (Walters et al. 2017). This difference leads to some large discrepancies in T2m that spread across adjacent regions throughout the forecast (e.g., Russia; Fig. 2d). Future work will align the treatment of water bodies in CPLDNWP and UNCPLD configurations (e.g., use of OSTIA lake SSTs instead of NEMO where the former are available).

We reiterate that the observed degradation at short lead times of T2m in CPLDNWP compared to UNCPLD reflects an inconsistency between uncoupled analyses and coupled forecasts. UNCPLD does not suffer from this inconsistency, giving it an advantage in verification. With that disadvantage in mind, extensive areas with improved RMSE in CPLDNWP at longer lead times (blue in Figs. 2b and 2d) are encouraging evidence of additional skill from coupling. We expect these regions to verify even better in a coupled DA system.

c. SST forecast skill

In this section we assess the ability of the coupled NWP system to predict daily mean SST. As mentioned at the beginning of section 3 it is relevant within the context of this study to compare forecast skill of predicted SST to skill of persisted SST. We first verify predicted daily mean SST in CPLDNWP, out to day 7. Then we do the same for persistence. As a persistence forecast we use SST from the preceding day (i.e., the most recent SST available at each initialization time of 0000 UTC). Verification of both these forecasts uses FOAM-day-m2 analyses: the first day of the second DA cycle of the FOAM operational ocean forecasting system (Waters et al. 2014).

Day-2 and day-7 RMSE of CPLDNWP are shown in Fig. 3. By day 2 RMSE is less than 0.3 K over most of the ocean (Fig. 3a). It grows to around 0.5 K by day 7 over most of the oceans (Fig. 3b). Notable exceptions with worse RMSE are regions of strong baroclinic instability: by day 2 RMSE in the Southern Ocean is 0.5 K, while in some other currents (Gulf Stream, Kuroshio, and Agulhas retroflection) it is already 1 K. By day 7 RMSE in these areas can exceed 2 K. This distinction between regions of “good” and “poor” forecast quality is also seen in bias (Figs. 3c,d): large tracts of the subtropical oceans have a day-7 bias less than 0.1 K, but biases in excess of 1 K exist in the baroclinic currents. By estimating the ocean’s Rossby radius of deformation Hallberg (2013) suggest resolutions of at least $1/8^\circ - 1/12^\circ$ are required before processes at the ocean mesoscale are starting to be resolved in these areas. The ocean model resolution of $1/4^\circ$ used here is insufficient to resolve ocean mesoscale variability at mid- to high latitudes. While this in itself is not a new conclusion (see e.g., Hewitt et al. 2017 and

Fig. 3. Forecast errors of 0.5-m ocean temperature (K) in CPLDNWP (a) RMSE at day 2, (b) RMSE at day 7, (c) bias at day 2, and (d) bias at day 7. Verification is against FOAM Verification period is 26 Sep 2018–25 Sep 2019.
references therein), the question we address here is if, for the purpose of atmospheric NWP, these imperfect predictions of SST still constitute an improvement over persisted initial SST fields as used in uncoupled forecasts.

The answer to this question is clear from Fig. 4. By day 2 the difference in MAE between CPLDNWP and persistence is less than 0.1 K (Fig. 4a). However, in the western North Pacific and North Atlantic the difference is between −0.25 and −0.1 K, which means that, already by day 2 MAE of persistence exceeds that of CPLDNWP. By day 7 these regions of superior MAE in CPLDNWP extend across most of the extratropical oceans (Fig. 4b). The main areas where MAE in CPLDNWP is larger than in the persistence forecast are the Southern Ocean, part of the North Atlantic Current, and around the Maritime Continent. Another area is the Falkland–Malvinas Current system in the South Atlantic, which may contribute to the degraded T2m forecast over southern South America (Fig. 2b).

For predicting SST variability coupled forecasts beat persistence from days 2–7 of the forecast, as shown by the extensive regions with negative $S$ (Figs. 4c,d). This include most of the midlatitude oceans, including the Kuroshio and Gulf Stream. At lower latitudes skill of coupled forecasts of SST over persistence is limited to coastal upwelling regions in the subtropics and regions with tropical instability waves (equatorial Atlantic and eastern Pacific). Skill is also present in currents along shelf edges in the Arctic.

One region in Fig. 4d with reduced skill compared to persistence is along the Antarctic Circumpolar Current. For this region we show area-average error growth against lead time in Fig. 5, using the error decomposition from the start of section 3. The RMSE in both forecasts (Fig. 5b) is dominated by the variance term, Fig. 5c. Initially persistence has a worse variance error, which in this region is dominated by errors in forecasting the ocean mesoscale field. Variance of the error in CPLDNWP grows slightly more rapidly and overtakes persistence by day 4. Rate of error growth decreases with lead time and CPLDNWP is only slightly worse by day 7. However, it is important to remember that these are spatial aggregates and that locally the difference between CPLDNWP and persistence in the Southern Ocean can be bigger, cf. Fig. 3d. In this eddying region the area average of MAE is smaller than the variance term, Fig. 5a, suggesting the greater challenge in this region is
predicting evolution of the eddy field, rather than maintaining mean SST.

d. Sea ice forecast skill

In polar regions the ocean–atmosphere–sea ice system can experience substantial coupled interactions at NWP time scales (e.g., Pellerin et al. 2004). In UNCPLD forecasts sea ice thickness and fraction are persisted, thereby eliminating such interactions. Similar as was done in the preceding section for SST, in this section we assess skill for sea ice forecasts from including coupled processes. We calculate forecast error for daily mean aggregate sea ice fraction and thickness (i.e., aggregating all thickness categories) and compare the error with that from a persistence forecast. For verification we use again sea ice fields from FOAM-day-m2 analyses. In FOAM the only sea ice observations that are assimilated are for sea ice area (“aice”). Direct assimilation of sea ice thickness (“hice”) observations poses various challenges (e.g., Blockley and Peterson 2018) and is not done operationally in FOAM. Instead analyzed sea ice thickness is determined in the NEMO-CICE model as part of the sea ice area assimilation scheme, as described by Peterson et al. (2015).

Area averages of forecast errors over the Arctic (Fig. 6) show that errors from CPLDNWP are always smaller than from persistence at all lead times. RMSE is dominated by the variance term. Variance of the forecast error in CPLDNWP is superior to that of persistence at all lead times. For example, RMSE from persistence at day 5 is similar to that of CPLDNWP by day 7. For thickness improvement in skill is even larger: RMSE from persistence by day 2 is already comparable to the day 6 error from CPLDNWP.

In the tropics the situation is very different: Z500 RMSE and variance difference fields have a distinct large-scale structure.
Some regions have a degradation of RMSE in CPLDNWP (e.g., the Indian Ocean and parts of the Andes), which may be related to degradation in near-surface temperature (cf. Fig. 2b) but we have not explored this further.

Differences in variance of Z500 are negative across most of the tropical belt, with local maxima over the eastern Pacific and the Atlantic. These regions coincide with regions of tropical instability wave (TIW) activity (Lyman et al. 2007). We have already seen that CPLDNWP is more skilful than persistence in predicting SST variability in these regions (Fig. 3d). The improved skill in Z500 could therefore at least partly be linked to increased skill in predicting SSTs here. Additionally, there are small-scale, isolated regions of improved Z500 across the subtropics that are associated with tropical cyclone (TC) activity. This is a consequence of reduced TC central pressure bias in CPLDNWP that will be discussed in section 4b. Because variance of Z500 is relatively small in the tropics, using it for normalization leads to large relative changes here. A better diagnostic for the tropics is the velocity potential $x$, as it relates to the dynamically important divergent part of the circulation (see Fig. S3 in the supplemental material). We find that CPLDNWP has generally smaller forecast error for $x$ compared to UNCPLD, both at 500 and 250 hPa (Fig. S2). Improvements in the tropics are seen in all three annual cycles, which points to a robust improvement in CPLDNWP (Table S1).

U250 forecast errors also have the largest improvements in the tropics, Figs. 8c and 8d. Regions with the largest improvements (between 10% and 15%) are mainly over the ocean: the TIW regions again, but also the Indian Ocean and South China Sea. To see if this improvement in U250 is due to individual disturbances we show time series of the absolute error in CPLD and UNCPLD for 2018–19, averaged over the tropical belt, Fig. 9a. Most days U250 in CPLDNWP is superior to UNCPLD and the running sum of the difference (green curve in Fig. 9a) is negative throughout and decreases monotonically. This suggests that upper-level wind variability in the tropics are consistently better predicted in CPLDNWP during the whole year. This tropical wind example provides a very clear case where we can see the benefit of predicting SST. A similar improvement in U250 is seen in all three years (Table 2, 2nd row) and we consider it to be a robust result in the UM. These very clear improvements in CPLDNWP at $T + 168$ in the tropical mid- to upper troposphere ($U250$, $x_{250}$, and $x_{500}$) occur even though performance for T2m is more mixed (Fig. 2b). As we saw in section 3b, verification of T2m against OSTIA-based analyses puts CPLDNWP at a disadvantage, potentially masking the benefits of predicting SST. To explore the improved tropical performance in CPLDNWP we have verified precipitation forecasts against GPM observations (Huffman et al. 2019). Precipitation anomalies are strongly
coupled to dynamics of the tropical atmosphere at time scales from days to weeks, and verification against GPM observations are not biased toward either CPLDNWP or UNCPLD. Throughout most of 2018–19, tropical rainfall by day 7 is predicted better in CPLDNWP than in UNCPLD (Fig. 9b). The improved predictions of the tropical atmosphere are therefore, at least in part, likely to be caused by better predictions of tropical rainfall, which benefit from interactive air–sea coupling. The spatial pattern of the reduction in MAE for rainfall shows improvements in CPLDNWP over tropical oceans as well as over landmasses [e.g., South America, equatorial Africa, and Southeast Asia (Fig. S3)].

From this section and sections 3a and 3b it is clear that short- to medium-range forecasts for the atmosphere benefit, in certain regions, from interactive coupling to the ocean. It is conceivable that this is the result from representing the annual cycle of SST in CPLDNWP, which is excluded from UNCPLD (cf. section 2). We have investigated this possibility in sensitivity tests with low-resolution (N216, 0.5° × 0.8° atmosphere resolution) forecasts with CPLDNWP and UNCPLD. Here UNCPLD uses SST forcing that consists of a persisted anomaly added to a daily varying SST climatology, thus including the annual cycle in SST forcing. Improvements to Z500 and U250 in CPLDNWP relative to UNCPLD are broadly similar to Fig. 8 (not shown). Notwithstanding the difference in resolution, this suggests that the improvements in CPLDNWP that we have described are not dominated by accounting for the annual cycle of SST, but, rather, by the full spectrum of SST variability that affects air–sea interaction during the forecast.

b. Tropical cyclones

In this section we compare the tropical cyclone (TC) forecast performance in CPLD and UNCPLD configurations of the UM. This is investigated here by exploiting the long parallel dataset of coupled and uncoupled forecasts. We combine all available N1280 forecasts at the time of writing, between 11 July 2017 and 1 November 2019, which includes the highly active 2017 Atlantic hurricane season. We use the TC tracking algorithm developed by Heming (2017) to track named storms and verify forecasts of storm position and intensity against operational advisories. Over this period there occurred 239 named storms, with 1036–148 storm forecast cases depending on lead time. Tropical depressions are excluded. We recall from section 2 that between July 2017–September 2018 the CPLDNWP configuration uses 85 vertical levels, as opposed to 70 levels in UNCPLD. These extra 15 levels are mostly in the

![Fig. 7. Skill $S$ of sea ice forecasts from CPLDNWP by day 7 relative to persistence [cf. Eq. (6)]: (a),(c) sea ice fraction (nondimensional) and (b),(d) aggregate thickness ($m^2$). Verification is against FOAM (day-m2) for the period 29 Sep 2018–28 Sep 2019.](image)
stratosphere and we anticipate therefore that this difference has only limited effect on tropospheric processes at NWP time scales. Separate tests confirm that for TC forecasts there is no systematic difference between L70 and L85 CPLDNWP configurations (see section SM4.1).

Median TC central pressure bias is negative in UNCPLD forecasts from around $T_{148}$, which grows to $-12$ hPa by $T_{168}$, Fig. 10a. This means there is a tendency to overdeepen central pressure in UNCPLD. This bias has become more prominent in the UM with increased horizontal resolution and the new dynamical core (Heming 2017; Walters et al. 2017). In contrast, we find that CPLDNWP forecasts exhibit far less of a deepening tendency through the forecast, with the median bias amounting to $-2$ hPa by $T_{168}$. The more extreme cases of overdeepening in CPLDNWP are shown by the 5th percentile of pressure bias in Fig. 10a. In UNCPLD this reaches values in excess of 40 hPa by $T_{108}$, in CPLDNWP they are about half that for the same lead time. In fact, toward the end of the forecast the 5th percentile of CPLDNWP is as large as the 25th percentile of UNCPLD, showing the heavy negative tail in pressure error in UNCPLD is greatly improved in CPLDNWP. In contrast, the 95th percentile upper tail (i.e., cases where low values of TC central pressure are considerably under-predicted), is very similar in both configurations.

The cold SSTs that are formed in the wake of TCs (e.g., Yablonsky et al. 2015) and within its inner core (Cione and Uhlhorn 2003) can reduce surface latent heat flux around TCs and weaken its intensity. This cold wake is predicted by the CPLDNWP configuration, unlike in UNCPLD predictions where persistence of the initial SST precludes any feedback between SST and storm intensity. It is encouraging to see that, in the coupled configuration used here, interactive air–sea coupling has approximately the required strength to effectively eliminate TC overdeepening in the UM in 50% of the forecasts (Fig. 10a). We also find a small but systematic reduction in track forecast error (TFE) in CPLDNWP, compared to UNCPLD (Fig. 10b, note logarithmic scale). For the median value this reduction in TFE appears from lead times $> T_{114}$ and is seen for most percentiles in the second half of the forecast. In other words, improved TFE is not just seen in forecasts with very large errors (hundreds of kilometers), but also in forecasts with TFE that is already comparatively small. There is a systematic shift in the distribution of TFE in.

<table>
<thead>
<tr>
<th>Cycles 16–17, 17–18, 18–19</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Latitude range</td>
<td>ΔVar Z500 (%)</td>
<td>ΔVar U250 (%)</td>
</tr>
<tr>
<td>35°–60°N</td>
<td>0.1, −0.7, −1.5</td>
<td>−0.3, −0.5, −0.6</td>
</tr>
<tr>
<td>23°N–23°S</td>
<td>−3.1, −3.3, −6.2</td>
<td>−3.5, −4.0, −3.8</td>
</tr>
<tr>
<td>35°–60°S</td>
<td>−0.4, 0.2, 0</td>
<td>−0.1, −0.2, −0.3</td>
</tr>
</tbody>
</table>
CPLDNWP toward lower values for these longer lead times. At shorter lead times (\(< T + 84\)) track forecast errors are very similar in both forecast systems.

These improvements in aggregate track forecast error point to an average superiority of coupled forecasts. It is useful to understand how this superiority is achieved, as it may point to key processes that could act as a focus for further model development. We want to determine, for example, if the reduction in TFE in CPLDNWP is the result from a few individual storms or a small, systematic improvement in all TC cases. For this we need to disaggregate the TFE metric. We define three categories of “superiority of track error,” based on the difference in TFE_i at lead time t_i:

- \((\text{TFE}_{i,\text{coupled}} - \text{TFE}_{i,\text{uncpld}}) < 50 \text{ km}: \text{CPLDNWP superior (“CS”);}\)
- \((\text{TFE}_{i,\text{coupled}} - \text{TFE}_{i,\text{uncpld}}) > 50 \text{ km}: \text{UNCPLD superior (“US”);}\)
- \(|\text{TFE}_{i,\text{coupled}} - \text{TFE}_{i,\text{uncpld}}| < 50 \text{ km}: \text{both similar (“SIM”).}\)

We determine the number of individual forecasts of named TCs in each of the three categories, as a function of lead time. These numbers are expressed as the fraction of the total number of TC cases at each lead time in Fig. 11a (i.e., the sum of the three curves is 1 at all lead times). Up to \(T + 36\) all TC forecasts are in the SIM category (purple line). From \(T + 36\) the fraction of SIM cases starts to drop, with similar increases in CS (orange) and US (blue) categories. Random fluctuations obviously lead to a growing number of cases where forecast tracks in both systems start to diverge. Between \(T + 36\) and \(T + 84\) CS and US categories are equally occupied, indicating that neither forecasting system is superior. However, between

**Fig. 9.** Time series of absolute error of (a) U250 at \(T + 168\) and (b) day 7 mean precipitation (i.e., averaged between \(T + 144\) and \(T + 168\)). Errors are averaged over the tropics (23°S–23°N) for CPLDNWP (orange) and UNCPLD (blue) on the left ordinate. Shown in green is the running cumulative difference (CPLDNWP – UNCPLD) on the right ordinate.

**Fig. 10.** (a) Pressure bias (hPa) and (b) track error (km) as a function of lead time for CPLDNWP (orange) and UNCPLD (blue) configurations. Lines show 5th, 25th, 50th, 75th, and 95th percentiles (labeled on the left). Shading shows 95% confidence interval for each percentile, estimated from bootstrap resampling the data for each lead time 10000 times. The number of storm predictions included in each case is shown at the top.
$T + 90$ and $T + 132$ CS cases grow more rapidly than US. During this time-window interactive coupling is increasingly benefiting TC TFE. After $T + 132$ the fraction of CS cases is levelling off and growing at a similar rate as US. This motivates us to focus on differences in TC forecasts during the time window between $T + 90$ and $T + 132$.

The spatial distribution of CS and US cases for this forecast time window (Fig. 11b) shows that both these categories occur in all main TC regions. For example, we observe US cases in the northwest Pacific basin, where Feng et al. (2019) have documented a dominance in the UM of CS cases—albeit in a smaller sample size than used here. To determine if there are differences in the structure of the TCs and their surroundings we apply a compositing technique, averaging forecasts for CS cases (both by CPLDNWP and separately for UNCPLD) and similarly for US cases. To account for the difference in TC translation direction we apply a coordinate transformation to a TC-centered spherical coordinate system: the TC center is located in the origin and a rotation around the origin is applied so that the storm travels in negative pseudolatitude direction. This transformation is described in section S4.2. We restrict our compositing to Northern Hemisphere cases only, to avoid confounding signals that depend on the sense of circulation around the storm center. Composites for MSLP are shown in Fig. 12. TCs in CS cases (left column) are larger and deeper than US cases (central column), and the difference is shown in right column (Figs. 12c,f). We conclude that, when TFE in CPLDNWP is better than in UNCPLD, this occurs on average for larger, deeper storms. Conversely, when UNCPLD forecasts have better TFE than CPLDNWP, TCs tend to be smaller. Core pressure in CS cases is substantially higher (less deep) in CPLDNWP than in UNCPLD, by as much as 5 hPa (Fig. 12g), decreasing with radial distance. In US cases this difference between CPLDNWP and UNCPLD is reduced (Fig. 12h) and diminishes more rapidly with radial distance. We interpret Fig. 12g to suggest that overdeepening in UNCPLD is more pronounced in CS cases than in US cases, and that air–sea coupling in CPLDNWP alleviates this problem in CS cases. Differences between CS and US cases extend well beyond the storm center itself, with implications for the steering flow (e.g., upstream of the TC center) (see Fig. S5). Further analysis of air–sea interaction near TCs will be presented in future work.

c. Madden–Julian oscillation

To quantify the sensitivity of the MJO to two-way coupling between ocean and atmosphere we have calculated error and skill statistics for forecasts in both configurations (CPLDNWP and UNCPLD). The MJO in the predictions is quantified using the method developed by Wheeler and Hendon (2004). Tropical intraseasonal variability is projected onto a two-dimensional
FIG. 12. Composite matrix of MSLP. (left) TC cases when track forecast in CPLD is superior (CS), (center) when UNCPLD is superior (US), and (right) shows difference field. Composites (a)–(c) from CPLD forecast, (d)–(f) from UNCPLD forecast, and (g)–(i) the difference. Panel titles in the difference field indicate what difference is plotted, for example, (c) shows the difference between the field in (a) minus the field in (b). Composites are calculated for Northern Hemisphere storms, with track superiority determined between \( T = 92 \) and \( T = 132 \), as in Fig. 11. Dotted lines indicate the coordinate axes of the TC-following coordinate system, with the storm traveling in negative x direction. The extent of the domain in degrees is stated in (h).
phase space, with time evolution given by two one-dimensional time series, RMM1 and RMM2. We calculate four error metrics for predicted RMM1 and RMM2 that are routinely used to evaluate MJO performance in initialized predictions (Matsueda and Endo 2011; Kim et al. 2018).

We have verified the MJO in periods comprising boreal winter when activity is usually at its peak, from 1 October to 1 May, for all three years 2016–19 (729 days in total). We only verify days when there is an active MJO in the observations. This happens on 479 days (i.e., around 66% of the time). In each of the three boreal winter seasons (2016–17 through 2018–19) there are only small differences in forecast accuracy between the two prediction systems, without systematic superiority of either model configuration. We show the average over the three seasons in Fig. 13. The only systematic difference evident from Fig. 13 is for the phase error between predicted and observed MJO evolution: in coupled predictions an active MJO propagates, on average, with a smaller phase error than in uncoupled predictions. The latter develop an increasingly slow propagation bias. It has been shown for intraseasonal time scales that allowing SSTs to respond to MJO variability promotes MJO propagation (De Mott et al. 2016). It appears from our results that when SSTs can respond to MJO forcing this also benefits MJO propagation at much shorter time scales, up to 7 days. Errors in phase transitions of the MJO have the potential to modulate the timing of MJO teleconnections to the extratropics. Given the time lags involved this is more likely to affect medium- to extended-range than short-range forecasts (Matthews et al. 2004; Cassou 2008).

**FIG. 13.** Error metrics of MJO predictions in CPLNWP (orange) and UNCPLD (blue) forecasts as a function of lead time. (a) Bivariate RMSE, (b) bivariate Pearson correlation, (c) amplitude error of (RMM1, RMM2), and (d) phase error in degrees in (RMM1, RMM2) plane. Note (d) are not errors in the number of the MJO phase (1–8) itself. Verification is against MJO monitoring data from www.bom.gov.au/climate/mjo for the period 1 Oct–31 May for the years 2016–19 for days with an active MJO.
Apart from MJO propagation, at the short lead times of NWP forecasts changes in SST thus appear to be of limited importance to predicting MJO evolution in our sample. This is consistent with existing studies (Woolnough et al. 2007) who found predicting SST to become beneficial to MJO forecast skill for lead times of 10 days or more. Prescribing persisted SST through day 7 is, on average, not a poor approximation for predicting this mode of tropical variability in our sample of MJO cases studied. That is not to say that in other MJO cases predicting SST or, more generally, the state of the upper ocean, can be beneficial for forecast skill [e.g., when oceanic conditions are conducive for MJO triggering (Webber et al. 2012) or propagation (Shelly et al. 2014)].

5. Discussion and conclusions

As part of its preoperational testing the Met Office have been running daily forecasts with a coupled ocean–atmosphere forecasting system since 2016. By maintaining a direct traceability to the science configuration of the operational uncoupled forecasting model we can directly compare skill in both systems, to quantify the impact of including air–sea interaction. Verification is done against operational analyses from the uncoupled model. Skill in the free troposphere in the tropics is improved by two-way air–sea coupling. At midlatitudes the effect is neutral with no robust signal in the three years considered. Improvements for TC track and pressure forecasts occur at longer lead times. They are most evident between $T + 90$ and $T + 132$ and for the largest storms. For smaller storms coupling can sometimes be detrimental. For the current atmospheric science configuration skill in MJO is largely independent of coupling, apart from a small impact on propagation error.

In most of the tropical and subtropical oceans SST variability predicted by the coupled model is at least as good or better than persisted SSTs used in uncoupled forecasts. Midlatitude ocean SST, particularly in Southern Ocean is less skillful than persistence. This is consistent with the marginal ability of the $1/4^\circ$ ocean model to resolve the Rossby radius of deformation at mid- to high latitudes. We anticipate that increasing ocean resolution to $1/12^\circ$ will improve midlatitude SST forecasts and work is currently underway to update the ocean model resolution, including that of the DA system. Another issue we have encountered is the treatment of shallow water bodies (small and large lakes, enclosed sea basins, shallow marginal seas). These are treated differently in coupled and uncoupled simulations, making it difficult to perform a meaningful comparison of forecast skill for the atmosphere in such regions. The effect of these differences can spread across subcontinental scales during a 7-day forecast. We hypothesize that the aforementioned problems in midlatitude oceans and shallow water bodies contribute to the lack of any additional skill in midlatitudes. We anticipate that planned model improvements in these areas will offer benefits to atmosphere forecasts at midlatitudes.

Acknowledgments. This work was funded by the Public Weather Service Contract. We thank our former colleagues Ann Shelly and Livia Thorpe for their contributions in developing the Met Office coupled NWP system. We thank Eun-Jung Kim for running the L70 CPLDNWP suite. Comments from two anonymous reviewers and Alistair Sellar have helped to improve this paper and are gratefully acknowledged. We thank Chris Harris, Rob Darvell, and Julian Heming for their helpful discussions and Andrew Hartley for providing the GPM data.

Data availability statement. The nature of the 4D data generated in running the model forecasts requires a large tape storage facility. Data volumes are of the order of 70 TB for one year of 6-day forecasts with the NEMO/CICE system, around 400 TB for one year of 7-day 0000 UTC forecasts with the global UM. NEMO/CICE forecast data and OSTIA SST analyses are available on the CEMES web portal http://marine.copernicus.eu For global UM and CPLDNWP forecasts used in this study there currently are no on-demand facilities to accommodate ad-hoc data requests. Anyone who is interested in accessing these data should consult the information provided on https://www.metoffice.gov.uk/services/data. GPM precipitation data were provided by the NASA/Goddard Space Flight Center via their website at https://gpm.nasa.gov/data-access.

REFERENCES


