The Benefits of Ensemble Prediction for Forecasting an Extreme Event: The Queensland Floods of February 2019

MATT HAWCROFT,^{a,b} SALLY LAVENDER,^{a,b} DAN COPSEY,^b SEAN MILTON,^b JOSÉ RODRÍGUEZ,^b
WARREN TENNANT,^b STUART WEBSTER,^b AND TIM COWAN^{a,c}

^a *University of Southern Queensland, Toowoomba, Australia*

^b *Met Office, Exeter, United Kingdom*

^c *Bureau of Meteorology, Melbourne, Australia*

(Manuscript received 9 October 2020, in final form 22 April 2021)

ABSTRACT: From late January to early February 2019, a quasi-stationary monsoon depression situated over northeast Australia caused devastating floods. During the first week of February, when the event had its greatest impact in northwest Queensland, record-breaking precipitation accumulations were observed in several locations, accompanied by strong winds, substantial cold maximum temperature anomalies, and related wind chill. In spite of the extreme nature of the event, the monthly rainfall outlook for February issued by Australia's Bureau of Meteorology on 31 January provided no indication of the event. In this study, we evaluate the dynamics of the event and assess how predictable it was across a suite of ensemble model forecasts using the Met Office numerical weather prediction (NWP) system, focusing on a 1-week lead time. In doing so, we demonstrate the skill of the NWP system in predicting the possibility of such an extreme event occurring. We further evaluate the benefits derived from running the ensemble prediction system at higher resolution than used operationally at the Met Office and with a fully coupled dynamical ocean. We show that the primary forecast errors are generated locally, with key sources of these errors including atmosphere–ocean coupling and a known bias associated with the behavior of the convection scheme around the coast. We note that a relatively low-resolution ensemble approach requires limited computing resources, yet has the capacity in this event to provide useful information to decision-makers with over a week's notice, beyond the duration of many operational deterministic forecasts.

KEYWORDS: Australia; Ensembles; Forecasting techniques; Numerical weather prediction/forecasting

1. Introduction

From late January to early February 2019, widespread flooding inundated large parts of northeast Australia associated with a quasi-stationary monsoon depression. The monsoon depression lasted for 10 days, with over 400 mm of precipitation falling over a large proportion of northwest Queensland and on Queensland's east coast, including the regional city of Townsville (population: 180 000), daily maximum temperatures 8°–10°C below normal, and wind gusts above 70 km h^{−1} (Cowan et al. 2019). For weather stations that have more than 30 years of observations, 4 set new daily rainfall records for any month of the year, 18 set new 7-day records for any month, and 21 set new 10-day records for any month (Bureau of Meteorology 2019b), indicating that it was at the near-weekly time scale for which this event was most extreme. Considerable losses were sustained in agricultural communities in northwest Queensland, with over 620 000 head

of cattle lost to a combination of drowning and hypothermia brought on by exposure to extreme rain and windchill associated with the monsoon depression. Peak inundation covered around 200 000 km² of the river catchments flowing into the Gulf of Carpentaria and the impact of the flooding on the productive capacity of the land will impact farming capability in some areas for multiple years (Hall et al. 2020). On the northeastern coast around Townsville, over 3000 homes were flooded. For northwest Queensland, the week from 31 January to 6 February represented the most extreme week of weather. The depression developed in the context of a delayed monsoon onset over northern Australia and record hot conditions from November 2018 to January 2019 (Bureau of Meteorology 2019a). This sudden switch from hot and dry to wet and cool reportedly contributed to the impact of the event as the condition of livestock was already weakened by the preceding conditions (M. Munchenberg 2019, personal communication).

Cowan et al. (2019) investigated the atmospheric conditions during the event and the event's predictability on the weekly to subseasonal range. They showed that at the time of the event, the tropical convective signal of the Madden–Julian oscillation (MJO) was over the western Pacific, and likely contributed to the heavy rainfall, but that neither the El Niño–Southern Oscillation (ENSO) or Indian Ocean dipole (IOD) were in the usual phases conducive to increased rainfall over Queensland. Off Australia's east coast, an anticyclone in the northern Tasman Sea helped to maintain a positive phase of the southern annular mode (SAM), which promoted onshore easterly flow. As such, they concluded that

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/MWR-D-20-0330.1.s1>

Corresponding author: Matt Hawcroft, Matthew.Hawcroft@usq.edu.au

DOI: 10.1175/MWR-D-20-0330.1

© 2021 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

the most important drivers of this event appeared to have been the MJO, the northward-shifted blocking anticyclone, and the SAM, but not ENSO or the IOD. Such a conclusion implies a relatively limited temporal range (days to weeks) of predictability of the event, since numerical models of the resolutions typically used for multiweek to seasonal prediction, which require convection to be parameterized, have a tendency to poorly capture the propagation of the MJO (Xiang et al. 2015).

The dynamical characteristics of the event were unusual in that depressions frequently pass into northern Queensland during the monsoon season but more typically move rapidly away from the region. This depression was quasi-stationary for several days and it is this atypical behavior that is the primary reason the observed weather conditions were so extreme. Cowan et al. (2019) evaluated experimental forecasts from the Australian Bureau of Meteorology's (BOM) ACCESS-S1 (Australian Community Climate and Earth-System Simulator–Seasonal Version 1) ensemble prediction system, with further analysis of ensemble forecasts from four other prediction systems within the Subseasonal to Seasonal (S2S) Project database (Vitart et al. 2017). Cowan et al. (2019) showed, using forecasts initialized on or around 24 January (1-week lead time), that the S2S models typically forecast the precipitation peaking too close to the northern coast relative to observations, with an underestimate of the inland precipitation accumulations that caused considerable flooding south of the Gulf of Carpentaria, over northwest Queensland. Cowan et al. (2019) did, however, show that individual members of each ensemble predict higher totals. In the ACCESS-S1 forecasts, the spread of the ensemble forecasts was such that it gave a doubling of the likelihood of extreme conditions (highest quintile) when referenced against a 1990–2012 observed climatology, which they concluded could be viewed as a successful forecast.

We note that ensemble forecasting is the recognized approach to deal with the inherent uncertainty of weather and climate forecasts (Buizza et al. 2005). Here we extend the work of Cowan et al. (2019) to further investigate the predictability of this event at the multiweek time scale, beyond the range of the detailed weather forecasts produced by many operational forecast centers. We emphasize that the purpose of such ensemble forecasts is to characterize the distribution of possible events, rather than provide forecasts of such absolute amounts, and are particularly useful to identify the likelihood of extreme events occurring, as in this event. Here we employ a suite of ensembles based around the Met Office's (UKMO) Unified Model (UM), as used operationally in the UKMO Global Seasonal forecast system, version 5 (GloSea5; MacLachlan et al. 2015) and in both the UKMO and BOM operational numerical weather prediction (NWP) systems. One benefit of this approach is that the atmospheric model component of ACCESS-S1 is the same as GloSea5 and is run at the same horizontal resolution of N216 (~60 km in the midlatitudes) with 85 vertical levels. This study therefore represents a natural extension of the work of (Cowan et al. 2019).

In this study, we ask a series of questions relevant to the multiweek predictability of this event. First, we evaluate the

ability of the ensemble to predict the likelihood of the extreme weather conditions which were observed in northwest Queensland, both in terms of capturing the precipitation and temperature extremes and also the observed dynamics which caused those extremes. As part of this, we evaluate whether additional skill is gained in the forecast ensemble by (i) increasing the resolution of the model, (ii) using a newer version of the model with improved physics and (iii) coupling the atmospheric model to a dynamical ocean. All of these choices employ additional computing resource and inform the ability of forecast centers to adopt multiweek ensemble forecast systems operationally.

To isolate the mechanisms which drive biases and errors in the forecasts, our analysis incorporates the use of an innovative nudged ensemble suite, where dynamical fields in forecasts are (regionally or globally) nudged back toward analysis, and a nested suite, where a higher-resolution model is embedded within the global model.

Nudging has been used in previous studies to investigate sources of model error in the UM (e.g., Johnson et al. 2019; Rodríguez and Milton 2019) and the novel aspect of the approach in this study is to employ that technique in an ensemble framework, to provide a more rigorous analysis of predictability and sources of model error. In a nudging framework, the circulation in the model is relaxed back toward some reference value, such as operational analysis, to remove large-scale circulation errors in forecasts. This nudging might be applied globally or in a regional domain, the latter allowing an investigation of the geographical source of forecast errors. Through the multiensemble approach applied here, we seek to identify where errors occur in forecasts for this specific event, but also seek to relate any such errors back to the overall performance of the ensemble prediction system in order to evaluate where forecast centers may utilize their limited computing resource to deliver the most useful information to end-users at multiweek time scales.

In the nested suite, a high-resolution (4.4 km) version of the UM is run over a limited domain within a global version of the UM. This high-resolution regional model is convection permitting. Parameterized convection is a well-known source of forecast biases in lower-resolution numerical models (e.g., Birch et al. 2015). Here, it is run in an ensemble framework for the first time. Table 1 provides a summary of the experiments evaluated herein.

This study begins by describing the data and methods used in section 2, before going on to evaluate and discuss the results from the ensemble in section 3. Conclusions are provided in section 4. A map is shown in Fig. 1 which denotes the locations discussed in the text and shows the orography of the region, with the Great Dividing Range being particularly relevant in the context of the precipitation distributions.

2. Data and methods

a. Model ensembles

The model ensembles used in this study are based on the UKMO UM (Hewitt et al. 2011). The baseline forecast

TABLE 1. Experiments analyzed in this study.

Name	Resolution	Ensemble members	Driving model (for nested suites)
GA6-N216	N216	25	—
GA6-N768	N768	25	—
GA8-N768	N768	25	—
Coupled	N768/0.25	25	—
Nudged	N216	25	—
Nested	4.4 km	25	GA6

ensemble is run at N216 (~60 km in midlatitudes), which is the resolution of the GloSea5 seasonal prediction system. Higher-resolution global forecasts are also produced at N768 (~17 km in midlatitudes). The physical version of the baseline model used here is Global Atmosphere, version 6 (GA6; Walters et al. 2017), which is the atmospheric configuration used in GloSea5 (MacLachlan et al. 2015) and was the operational forecast model in January/February 2019. For comparison, the most recent science configuration, GA8, is also used, which includes improvements to the convection, large-scale precipitation and boundary layer schemes. These forecasts are produced with sea surface temperatures (SSTs) that are persisted from the initialization time of the forecast, which, until relatively recently, has been standard in operational weather prediction. The basis for this approach is that at short time scales (a few days) it is typically felt that skill in predicting the weather is more likely to be increased by improving the capability of the

atmosphere model than employing an ocean model (Bauer et al. 2015; Brassington et al. 2015) and that the addition of a dynamic ocean adds considerable computing expense to the forecast system. As such, the SSTs from the initial state of the model are propagated forward in time. This has two primary impacts, first the broader pattern of the SSTs is unable to evolve with the weather (which becomes more important at longer lead times) and second, any higher-time-frequency coupled feedbacks between atmosphere and ocean are lost, which may influence the evolution of small scale phenomena, including tropical depressions and cyclones (Belcher et al. 2015; Smith et al. 2018; Vellinga et al. 2020) and the MJO (Vitart et al. 2017). A coupled ensemble is therefore also employed in this study which uses the same GA6 atmospheric physics at N768 (~17 km), coupled to dynamical ocean and sea ice models based on the same ocean and sea ice physics as was used in the Global Climate Version 3.0 climate model (GC3.0; Williams et al. 2018). The ocean component uses the NEMO (Madec et al. 2016) model using a 0.25° resolution grid (ORCA025). The sea ice component uses the CICE (Hunke et al. 2015) model using the same ORCA025 grid. The ocean and sea ice components communicate with the atmosphere every hour via the OASIS3-MCT coupler (Valcke et al. 2013). Ocean and sea ice fields are initialized from the Met Office operational ocean/sea ice analysis system, FOAM (Blockley et al. 2014). This coupled ensemble framework is based on an experimental deterministic coupled forecast system currently being tested at the Met Office. This framework has been

FIG. 1. Map of northern Queensland, showing orography (m) and the key towns and areas discussed in the text. The Great Dividing Range runs southward from the eastern coast of Cape York. Orography data from [GEBCO Compilation Group \(2020\)](#).

developed as part of the Met Office's plan for operational coupled NWP implementation by 2022, the benefits of which are described in [Vellinga et al. \(2020\)](#).

An ensemble prediction scheme aims to use multiple forecasts to simulate the evolution of the uncertainties at all stages in the prediction, and throughout the domain of a prediction model. One of the uses of ensemble forecast spread is to estimate forecast uncertainty and compute probabilities of forecast fields exceeding thresholds of interest. For operational weather forecasting purposes at the time of this work, the 44-member Met Office Global and Regional Ensemble Prediction System (MOGREPS; [Bowler et al. 2008](#)) was run four times daily at 0000, 0600, 1200, and 1800 UTC. Due to affordability reasons, only a subset of 18 members at each time are run to produce 7-day forecasts, providing a total of 36 members across two adjacent start times. MOGREPS is primarily designed to aid forecasting of high-impact weather systems, hence the focus on fewer members at higher resolution. Operationally, it is run at N640 (~20 km), providing additional information on predictability and probability to that from the higher-resolution global deterministic forecast model, which is run for 6 days at N1280 (~10 km). The initial conditions are generated through the combination of an analysis produced by the four-dimensional variational (4DVar) data assimilation ([Rawlins et al. 2007](#)) with perturbations generated by an ensemble transform Kalman filter (ETKF; [Bishop et al. 2001](#)). The ETKF generates ensemble perturbations representing the magnitude and statistical error structure of the uncertainties in the initial conditions.

In this work, we use the initial conditions of the first 25 members from the 0000 UTC cycles of the daily MOGREPS ensemble as the starting conditions for all simulations. The subset of members runs the risk of not being fully centered on the original 44-member ensemble set and may not sample all of the variability, but we do not believe this to be significant, as each member is an independent realization of the uncertainty. An offline comparison of a 25-member and 36-member ensemble showed that the ensemble means only changed slightly, and were largely equivalent. It would be too complex to pre-select which members to run as full-length forecasts based on spread at the initial time because perturbations in each member would grow at different rates during the forecast. However, it is possible that a somewhat higher variance could be achieved with optimal member selection.

In the operational MOGREPS ensemble, SSTs (and the model physics) are perturbed. Here, we initialize all atmosphere-only model ensembles with unperturbed analyzed SSTs in order to provide a like-for-like comparison between the atmosphere only and coupled forecasts (where the ocean is initialized from analysis across the ensemble). The model is run without perturbations to the physics. We note that an analysis of the skill of the atmosphere-only forecasts to predict the observed precipitation and temperature extremes for this particular event was not statistically different if perturbed SSTs were used (not shown).

b. Nudged experiments

To explore the potential role of remotely sourced errors on the predictability of the event, we have also carried out a series

of atmospheric nudging experiments using the same GA6-N216 atmosphere-only configuration of the UM evaluated elsewhere in this study. In these simulations, the free-running model simulation (Control) is relaxed back to the Met Office operational analysis over the globe (Global) or subdomains where there might be significant systematic errors impacting the predictability of the event through remote forcing of the circulation in northern Queensland. As identified in the study of [Cowan et al. \(2019\)](#), the most likely source regions for important teleconnections on the time scale of these forecasts are the Southern Ocean, the Tasman and Coral Seas, and a tropical band running through the Maritime Continent and western Pacific. These three regions encompass observed Southern Ocean jet variability during the event, a blocking high to the east of continental Australia and the active MJO event which was passing through the Maritime Continent and western Pacific, respectively. Simulations are therefore performed with regional nudging covering each of these domains to assess whether correcting any systematic biases in these regions has an impact on the predictability of the event. The simulations are nudged back to UKMO UM analysis temperatures and winds with a 6-hourly relaxation time scale. We do not nudge moisture, in part because it is technically difficult in this model, but also because it would impact the behavior of physics schemes (e.g., convection) which interact with and alter the environment in which they operate. In addition, for the regional nudging experiments, a 10° buffer zone around the relaxation subdomain is introduced in which the nudging increments are exponentially damped to zero to ensure a smooth transition between the nudged and free-running parts of the simulation.

c. Nested experiments

It has been shown that using kilometer-scale models for NWP can improve the location, intensity, and timing of forecast convective weather events (e.g., [Clark et al. 2016](#)). At horizontal grid lengths smaller than 5 km, NWP models can begin to resolve the largest scales of individual convective clouds and, as such, are referred to as convection permitting models (CPMs). CPMs can improve both the physical realism and the skill of rainfall forecasts when analyzed on appropriate scales ([Roberts and Lean 2008](#); [Schwartz and Sobash 2017](#)). CPMs have also been shown to improve the representation of the diurnal cycle of convection compared to models with parameterized convection ([Kendon et al. 2012](#); [Birch et al. 2015](#)). Running a CPM over a global domain incurs significant computational expense, making it impractical for many forecasting purposes. They are therefore frequently employed in nested suites over a limited domain, with the boundary conditions at the edge of that domain driven by a global (lower-resolution) forecast. Here, in order to assess the impact of removing any locally sourced biases from parameterized processes, such as the convection scheme, we run nested suites with a 4.4 km regional model (RA2; with an earlier version of the model described in [Bush et al. 2020](#)) with a 20° × 20° domain, centered on northwest Queensland (20°S, 140°E). In these experiments, we produce an ensemble of nested forecasts using the UM for the first time, allowing a more robust evaluation of the benefits of an embedded regional model than in any previous work. We

FIG. 2. Observed precipitation accumulations (mm) for the week of 31 Jan–6 Feb 2019, inclusive. Data are shown for (a) AWAP, (b) the mean (“GPI”) of the satellite-derived datasets shown in (d)–(f), (c) AWAP over land and the mean of the satellite derived datasets over ocean, (d) IMERG, (e) GPCP, and (f) PERSIANN. The + shows the location of the town of Normanton and the × shows the location of the town of Winton. The white line in (a) runs (left to right) between the towns of Cloncurry and Richmond. AWAP is only reliable over land so is not shown over ocean points.

use the same MOGREPS initial conditions as used elsewhere in this study, with the driving boundary conditions for the duration of the subsequent forecast from the N216 global model. Boundary conditions are updated 6-hourly and the runs are atmosphere-only.

d. Observations

Observed daily precipitation and temperature data over terrestrial Australia are taken from the BOM’s 5 km resolution Australian Water Availability Project (AWAP) gridded datasets (Jones et al. 2009). The AWAP data are generated using a weighted averaging process from station data and are not reliable over the ocean. To constrain precipitation accumulations on ocean points, satellite-based estimates are therefore required. We evaluate precipitation estimates from three datasets that are frequently used for analysis of tropical and subtropical precipitation, the Integrated Multi satellite Retrievals for GPM (IMERG; Hou et al. 2014) V06B data, the Global Precipitation Climatology Project 1DD v1.3 dataset (GPCP; Huffman et al. 2001) and the Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks dataset (PERSIANN; Sorooshian et al. 2000). We note that all satellite-based estimates are subject to some uncertainty and error (e.g., Rauniyar et al. 2017). In the present study, the AWAP data validates to gauge data in many locations. In Fig. 2 we show accumulated precipitation from 31 January 2019 to 6 February 2019 (inclusive), being the

period most relevant to the flooding in northwest Queensland identified by Cowan et al. (2019). The precipitation estimates from the satellite-based datasets span the AWAP data in the regions where the highest accumulations are recorded by AWAP. The satellite estimates include peaks around the town of Winton, to the south, and the town of Normanton and the western coast of Cape York, to the north of this area. The observed accumulations for this week were 142 mm in Winton and 247 mm in Normanton such that the satellite observations also span the gauge estimates. We therefore use AWAP precipitation data over land. We note that the mean of the three satellite products (“GPI,” being GPCP, PERSIANN, and IMERG) verifies more closely with the AWAP data over land than any of the three individual products. Over the ocean, there is no easy way to discriminate between the performance of the three satellite-based estimates and given the closer correspondence the mean of these products has with AWAP over land compared to any individual product, we use the mean, GPI, in our analysis for ocean points (Figs. 2b and 2c). Where dynamical fields are used in this study, they are from the Met Office operational analysis.

3. Results

a. Observed event dynamics

The observed dynamical evolution of this event is described in detail in Cowan et al. (2019) and a summary of the key

FIG. 3. Daily observed precipitation (mm) and daily mean sea level pressure (hPa) from the Met Office analysis for 23 Jan–6 Feb 2019. Precipitation on land points is the AWAP data and over ocean points is the mean of GPCP, PERSIANN, and IMERG (GPI). The core pressure of the low is shown as a black + with the pressure annotated. Contours are even numbers, plotted every 2 hPa.

features is provided here. The low pressure system developed over Cape York during 23–26 January (see Fig. 3), where it merged into a shallower monsoon trough before the center of the low strengthened and tracked eastward from 31 January to 1 February, with the core of the low then centered on Queensland's Gulf Country. The low then stalled for several days before gradually moving eastward between 5 and 6 February. During the period where it was quasi-stationary over the Gulf Country, it drew in moist air from the Coral Sea and Gulf of Carpentaria. It was the stalling of the depression and its interaction with the wider atmospheric conditions that resulted in the flooding event in northwest Queensland. Of note is the evolution of the event from 30 January to 2 February, as a broad monsoon trough formed across the tropical north of Queensland before deepening and moving east. Around this time, an anomalous anticyclonic feature was situated between the Coral and Tasman Seas to the southeast, providing onshore easterly flow to northeast Queensland. This moisture source combined with the quasi-stationary depression appear to be the key regional dynamical features that made this event so extreme.

In the peak week of 31 January to 6 February, over 400 mm of rain fell over a large proportion of northwest Queensland (Figs. 3 and 4a) and along the northeast coast, with a band of peak accumulations exceeding 600 mm stretching between the northwest Queensland towns of Cloncurry and Richmond (a distance of almost 300 km, see Fig. 2a) within catchments that

feed into the Gulf of Carpentaria. On the southern flank of the highest precipitation accumulations, the most extreme cold temperature anomalies were experienced (Fig. 4c), with daily maximum temperature anomalies as low as -10°C (for detailed discussion see Cowan et al. 2019). The spatial and temporal variability of the rainfall associated with the event is highlighted in Figs. 3 and 4a,b), where the topographical separation of coastal rainfall (e.g., Townsville) to the inland Gulf stations west of the Great Dividing Range can be observed. Figure 4b shows time series of precipitation accumulations covering a "Peak" region, capturing the most extreme inland precipitation, a larger "Gulf" region which covers much of the Flinders/Normanton river catchment areas and a "North Queensland" domain, which covers the wider region including around Townsville and the east coast, where substantial precipitation was also observed. In analysis throughout this study, these domains only include precipitation falling over land. Figure 4b shows the accumulation of precipitation in these domains in the 2 weeks from 25 January, evidencing the considerable increase in precipitation in the Peak/Gulf regions from 31 January onward. The large accumulations on the coast around Townsville from 25 January onward (see Figs. 2 and 3) account for the more linear accumulation of precipitation over the larger North Queensland domain during the 2 weeks shown.

An important initial question, already discussed in some detail by Cowan et al. (2019), is to ask how significant these

FIG. 4. Observed precipitation accumulations (mm) and mean surface air temperature (°C) from the AWAP data for (a) the total precipitation from 31 Jan to 6 Feb 2019; (b) precipitation accumulations from 25 Jan to 6 Feb for the Peak (blue) precipitation, the Gulf Country (red), and North Queensland (green) regions [the three domains shown in (a) and (c), with Peak being the smallest and NQld the largest]; (c) the mean temperatures from 31 Jan to 6 Feb; and (d) daily mean temperatures for the three regions, as in (b).

totals are in a historical context. Figure 5, column (c) (black bars), shows the AWAP observed accumulations for the week 31 January–6 February for the three regions. In Fig. 5, column (a), boxplots of historical observations (weekly accumulations for the DJF season from December 2000–December 2018, totaling 1710 samples) are shown for the three domains. In all three domains, the observed accumulations were considerably greater than any previously recorded amounts, almost double in the case of the Peak region (which may be partly accounted for based on the highly selective nature of the domain). The extreme weather during this event included cold maximum temperature anomalies (Fig. 4c and 4d) and we note that the temperatures from 31 January to 6 February were also unprecedented compared to historical DJF mean weekly surface air temperature observations (Supplementary Fig. 1a in the online supplemental material) for all three domains, with the Peak region almost 10°C below the climatological weekly mean for this season. It is therefore evident that this event was, on the temporal and spatial scales evaluated here, of unprecedented magnitude and had considerable impact on the communities and farming businesses in the areas worst affected (Cowan et al. 2019).

Most NWP forecasts are run for up to a week. With progressive improvement in forecast skill, many forecast centers, including the UKMO and BOM, are using seasonal prediction systems at multiweek time scales. Here we evaluate the ability of the UKMO's UM to provide useful forecast information

into the second week of the forecasts. This time period coincides with the likely limits of predictability for this event discussed in the Cowan et al. (2019) study and the skill indicated by an initial evaluation of operational GloSea5 forecasts undertaken as a precursor to this study (not shown). The provision of information to the public and key stakeholders at this lead time would allow for an additional degree of preparedness which was not within the capability of the existing forecasting and information dissemination infrastructure at the BOM at the time of the event.

b. Ensemble forecast skill

Ensemble forecasts were initialized at 0000 UTC 25 January, such that the week where much of the precipitation occurred (31 January–6 February) is week 2 of the forecast. In Fig. 5, columns (d)–(g), ensemble forecast accumulations from the GA6-N216, GA6-N768, GA8-N768, and coupled simulations are shown for the three regions. Bars show the ensemble mean, with vertical black lines showing the 5%–95% confidence interval on the mean (calculated using the *t*-distribution). Each ensemble member is shown as a black dot. Figure 5, columns (a) and (c), demonstrate that the observed weekly accumulations are record breaking for all three regions. Before discussing the detail of the forecasts, a further question that arises is whether these forecasts are extreme in the context of the model's own climatology, which will inevitably have some biases compared to observations. It is not possible to run a

FIG. 5. The 7-day precipitation accumulations (mm) for the (top) Peak, (middle) Gulf, and (bottom) North Queensland regions. Columns (a) and (b) are box-and-whisker plots showing the climatological distribution of 7-day DJF precipitation totals in AWAP observations (AWAPClim) and MOGREPS (MogClim) ensemble forecasts. Open circles show individual outlying data points. Column (c) is the observed AWAP accumulation from 31 Jan to 6 Feb. Columns (d)–(l) show precipitation accumulations for the 31–6 as the ensemble mean (bars), 5%–95% confidence intervals on the mean (vertical lines), and individual ensemble member forecasts (dots) for ensemble forecasts discussed in the text. Columns (d)–(f) are the uncoupled global forecasts, with (g) showing the coupled N768 ensemble forecast. Columns (h)–(l) show nudged forecasts, being the Control (free-running) simulation, the Global nudged simulation, and forecasts where the Southern Ocean, Coral and Tasman Seas, and Maritime Continent are nudged, as discussed in the text. Precipitation is shown for land points only.

climatology of all these individual ensemble forecasts due to the computational cost of such an endeavor. We therefore provide some context by evaluating a climatology of forecasts in the MOGREPS ensemble, which is run at N640 (~20 km). In Fig. 5, column (b), a boxplot of 8818 individual DJF forecasts from 2015 to 2018 is shown. The accumulations are weekly (week 1) accumulations from the 36 daily MOGREPS operational forecasts. MOGREPS was not run out to 7 days until 2015, so it is not possible to calculate these statistics for a longer or earlier period. It is evident that the MOGREPS ensemble is positively skewed compared to the AWAP observations (see Supplementary Fig. 2). For the Peak region, the observed (AWAP) 99th- and 99.9th-percentile events are 191 and 239 mm. In MOGREPS, the 99th and 99.9th percentiles are 264 and 428 mm. An evaluation of the MOGREPS forecasts above the 99th percentile for the Peak region shows that they are split into three groups. Two forecasts are apparently random single ensemble members on two different dates; 35 are from a cluster of forecasts in late December 2015 and 51 are from early March 2018. These two clusters both represent forecasts of observed precipitation accumulations which were, locally, in excess of 200 mm and above the 97.5th percentile for the Peak region observed climatology (not shown). As such, within MOGREPS the likelihood of a single forecast for the Peak region above the MOGREPS 99th percentile coinciding with no occurrence of an observed extreme event is less than 0.025% (based on the above sample of 2 of 8818 forecasts). We reiterate that the MOGREPS accumulations are a week 1 forecast and therefore not directly comparable to the current experiments, which are week 2 forecasts. We expect our experimental forecasts would be less skillful due to the additional lead time but in the absence of a climatology of week 2 ensemble forecasts are unable to provide such an evaluation of their skill.

In the 25 member ensemble forecasts shown in Fig. 5, columns (d)–(g), the four forecast ensembles all provide a strong indication of an event that is at the upper limits of the observed record and MOGREPS climatology for North Queensland. The ensemble means are well in excess of 100 mm and the distribution of forecasts includes several members indicating an event of unprecedented magnitude may occur. For the smaller Gulf region and the Peak region, the ensemble means are biased low relative to the observed accumulations, but all of the model ensembles include multiple members forecasting an event beyond the range of the historical observed record and above the 99th percentile in the MOGREPS climatology for both of these regions. In the Peak region, the N216-GA6 ensemble includes six forecasts over the 264 mm 99th percentile in MOGREPS, the N768-GA6 has six, the N768-GA8 has three, and the coupled ensemble has five. Such an occurrence has no precedent in the MOGREPS climatology and shows clear skill in the ability of the model to provide a skillful forecast into week 2. Cowan et al. (2019), evaluating ACCESS-S1 forecasts compared to historical station data for this event, found that the ACCESS-S1 ensemble spread was such that it gave a doubling of the likelihood of highest quintile (>80%) precipitation when referenced against a 1990–2012 climatology, which they concluded could be viewed as a successful

week 2 forecast. Here, we show that for this event, the model had the capacity to predict up to a 24-fold (6 members of a 25-member ensemble) likelihood of experiencing an event within the historical top 1%. For many livestock producers in northwest Queensland, the provision of such information at this lead time would have enabled response measures to be implemented that could have partly mitigated the impact of the event (M. Munchenberg 2019, personal communication). We note that in this experimental framework, we are unable to postprocess and bias correct the forecasts to reduce any systematic model errors. In an operational context, such an approach would improve the long-term performance of the ensemble in predicting the magnitude of extreme events, though for individual extreme events may not always improve the accuracy of those forecasts (e.g., Pantillon et al. 2018).

Given the significant difference in computational expense between the atmosphere-only simulations (particularly the N216 ensemble) and the coupled simulations, it is notable that Fig. 5 shows, that for total precipitation accumulations, there is no statistically significant difference in the ensemble mean of the forecasts (at 5%–95% confidence using the *t*-distribution). This suggests that no additional skill is acquired by using the higher-resolution model or coupled suite for this particular event, based on the total accumulated precipitation metric. This is somewhat surprising given it has been shown in other studies that there are clear physical benefits to running coupled forecasts, as, for example, air–sea fluxes are better represented (e.g., Belcher et al. 2015; Smith et al. 2018; Vellinga et al. 2020). In the current case study, the SSTs over the Gulf of Carpentaria were unseasonably warm prior to this event and cooled during the 2-week forecast period. This cooling was skillfully forecast by many of the coupled ensemble members (Supplementary Fig. 3a), both in terms of the mean cooling of around 1°C across the basin and the pattern of cooling where local cooling anomalies exceeded 2°C (not shown). Given the atmosphere only ensembles use SSTs from the 25 January analysis for the duration of their forecasts, any such changes in the pattern or temperature of the SSTs are not represented in those forecasts. This may therefore impact the dynamical evolution of the low pressure and the resultant daily precipitation and temperature anomalies. Changes in the mean SSTs in the Coral Sea were far lower during the 2-week analysis period (Supplementary Fig. 3b).

In Fig. 6, time series of precipitation accumulations are shown for 31 January–6 February for the three regions introduced above. While a forecast may have a reasonable total weekly accumulation, if the day-to-day accumulations do not compare well to observations, this implies that the dynamical evolution of the event was not well forecast. Figure 7 therefore also shows analyzed and ensemble-mean surface pressure for selected days during the forecast period, with the location and depth of the center of the depression from each ensemble member overlaid on each plot.

In Fig. 6, it can be seen that the ensemble-mean forecasts for all the model ensembles are considerably lower than the observed precipitation accumulations, with this bias increasing as the domain is reduced in size. We note that for an extreme event such as this, it might be unrealistic to expect the ensemble mean to accurately forecast the observed event. The

purpose of using an ensemble is to identify the likelihood of whether such an extreme event may occur—something that we would argue has been successfully achieved in these forecasts (see Fig. 5). For the North Queensland region, all of the ensembles include members which provide precipitation accumulations which follow the observed day-to-day accumulations, indicating multiple ensemble members which successfully forecast that the low pressure system remained within the wider North Queensland domain, resulting in substantial precipitation accumulations. The ensemble mean for the N768-GA8 configuration is lower than the three GA6 ensembles in all three domains (though this is not significant at the 5%–95% level, see Fig. 5), with fewer ensemble members that have high precipitation accumulations compared to the GA6 forecasts for both the Gulf and Peak regions. For the two smaller regions, Fig. 6 shows that while several ensemble members accurately forecast the total accumulations for the week, the day-to-day accumulations were not as well forecast. In several of the GA6-N216 and GA6-N768 forecasts, the precipitation accumulations are initially too high, before tailing off.

This behavior can be understood with reference to Fig. 7. In the first week of the forecast, there is some spread in the location of the low across all the ensembles, though up to 30 January, the low itself is a shallow trough where the core is not as well defined, such that the variability in the forecast location of the identified low center is greater than that of the wider trough structure itself (not shown). The operational analysis (see also Fig. 3) identifies the eastward migration and deepening of the low between 30 January and 1 February, during which period a number of the atmosphere only experiments erroneously forecast that the low moves too far to the east and northeast. During this period, several forecasts produce rapid accumulations of precipitation in the Peak and Gulf regions (Fig. 6) which then tail off as the low is forecast to migrate farther eastward too quickly. As such, several of the more accurate forecast weekly totals in the atmosphere only experiments are a case of “the right result for the wrong reasons” as the dynamics of the event were poorly forecast. In contrast, the coupled model ensemble has a more tightly clustered spatial distribution of surface pressure forecasts up to 3 February (forecast day 9) and a deeper ensemble-mean low. The analyzed low pressure then begins to move eastward and the spread in all the ensembles is large by 5 February (forecast day 11).

Figure 8 shows maps of total (31 January–6 February) and daily precipitation accumulations (every second day from 28 January) in observations and the model ensembles. There is a persistent underestimate of the total and daily precipitation accumulations around the Peak region across the ensembles as the model does not produce enough precipitation inland. There is also an erroneous persistence of precipitation along the western coast of Cape York in the three atmosphere only ensembles, leading to accumulations of over 600 mm (Figs. 8b–d) along this coast for the week of 31 January–6 February. This coastal precipitation bias is a known forecast error in the UM associated with an inadequate representation of sea-breeze dynamics, convergence, and the triggering of convection at the right time within the diurnal cycle (see Birch et al. 2015). We note that the

FIG. 6. Precipitation accumulations (mm) from 31 Jan to 6 Feb, inclusive for (a)–(e) the Peak, (f)–(j) Gulf Country, and (k)–(o) North Queensland regions. The black lines show the AWAP accumulations, with the individual ensemble members shown in the first four columns and the ensemble-mean forecasts in the right-hand column. Colors for the ensemble mean follow those for the individual ensembles. Precipitation is shown for land points only.

forecasts produced by the coupled model are more successful in forecasting the timing and distribution of precipitation over the southern Gulf of Carpentaria and western coast of Cape York. In the atmosphere only simulations, the SST patterns are persisted from the start of the forecast, whereas they evolve in the coupled simulation (Supplementary Fig. 3). These differences may manifest themselves via the pattern of the SSTs having an impact or the interaction between the dynamic ocean surface and the atmosphere improving the representation of surface fluxes.

This can be tested by running a series of experiments where the SST pattern in observations/analysis and the coupled model are used in atmosphere only runs as a specified boundary condition. In such experiments, the SST pattern evolves but any shorter time scale feedbacks from atmosphere to ocean are not represented. While the ensemble-mean precipitation forecasts for the three analysis regions have similar skill to those in Fig. 5 (not shown), Supplementary Figs. 4 and 5 show the precipitation forecasts associated with the experiments which have uncoupled (but

FIG. 7. Daily average mean sea level pressure (hPa) from the (left) analysis and (remaining columns) the model ensemble mean (as marked) for every second day within the period of the forecasts. The location of the core of the low pressure is shown for each ensemble member. The absolute amplitude of each low pressure corresponds with the scale shown. Contours are every 2 hPa.

time evolving) SSTs suffer from the same error as the atmosphere only forecasts with SSTs persisted from the start time, with fewer ensemble members predicting the stalling of the low to the south of the Gulf of Carpentaria. This suggests that small scale feedbacks between atmosphere and ocean (and their dynamical coevolution) play a significant role in explaining the difference in forecast skill between the

coupled and GA6/GA8 atmosphere-only ensembles evaluated in Figs. 5–8.

c. Sources of error: Local and remote processes

Residual error in the forecasts may be split into two sources for the purposes of further analysis. Errors may be sourced locally, in the region of the low pressure itself, or they may be

FIG. 8. (left) Observed and (remaining columns) ensemble-mean precipitation accumulations (mm) for (top) the total precipitation from 31 Jan to 6 Feb, inclusive, and (remaining rows) daily precipitation accumulations for the days shown. The scale for the total precipitation is in the top right of the figure; the scale for the daily accumulations is at the bottom of the figure.

sourced remotely and impact the evolution of the low through teleconnection patterns from their source region. In either case, the root source of the errors may be due to the poor representation of surface or atmospheric processes, including errors in the initialization of fields from analysis. To investigate these possible error sources in more detail we use three further tools—nudged and nested model ensemble simulations and a series of N216 ensemble forecasts initialized progressively closer to the start time of the event.

In the nudged simulations, the wind and temperature fields above the boundary layer in the model (either globally or in a specified domain) are relaxed back toward analysis fields. As such, any errors in circulation which are generated within the nudged region should be corrected. Precipitation and moisture are not nudged, and precipitation in nudged runs reflects the response of the parameterization to the “corrected” circulation and temperature fields. Where the model is nudged globally, residual errors in forecast fields can be attributed to

FIG. 9. (a) Observed and (b)–(f) ensemble-mean precipitation accumulations (mm) from 31 Jan to 6 Feb for nudged forecasts started on 25 Jan. The Control simulation in (b) is a free-running forecast with fixed SSTs. The Global simulation in (c) has winds and temperature relaxed to reanalysis above the boundary layer. The Southern Ocean, Coral Sea, and Maritime Continent simulations in (d)–(f), respectively, are only nudged within the specified region, with the remainder of the atmosphere evolving freely.

local parameterized processes, such as surface–atmosphere interactions, convection or the boundary layer. We note that in a free-running simulation, any errors in these small-scale processes may result in upscale feedbacks to the evolution of the wider circulation during the forecast, such that they may impact forecast skill. Within the framework of the nudging simulations, we are only able to ask whether any errors remain when the circulation is constrained and cannot use these simulations to rule out locally sourced errors as the root cause of forecast error in free-running forecasts. Where a specific region is nudged, teleconnection errors emanating from that region into northern Queensland should be substantially reduced. Cowan et al. (2019) identified likely source regions for error in the period of interest as the Southern Ocean (nudged region: 90° – 35° S, 0° – 360° E), the Tasman and Coral Seas (40° – 0° S, 165° – 190° E) and a tropical band running through the Maritime Continent and western Pacific (5° S– 10° N, 40° – 170° E). We therefore evaluate a series of forecasts run from the 25 January with the atmosphere nudged in these domains.

In Fig. 9 ensemble-mean precipitation forecasts for the period 31 January–6 February are shown (days 6–12). The baseline simulation, an N216 atmosphere only control forecast with no nudging (Fig. 9b) can first be compared to a simulation with global nudging (c). It is clear that where the model is run with global nudging of winds and temperature above the boundary layer, much of the forecast error is recovered. This implies that the model is capable of producing realistic

precipitation accumulations when the dynamics of the low are well forecast, suggesting that surface–atmosphere interactions and subgrid processes, such as convection and the boundary layer, are reasonably represented. We reiterate that small errors in these processes have the potential to have substantial impacts in free-running simulations as such errors may amplify and modify the evolution of the forecast dynamics. Turning to remotely sourced errors, forecasts where the Southern Ocean (Fig. 9d), Coral and Tasman Seas (Fig. 9e), and Maritime Continent (Fig. 9f) are nudged do not recover a substantial amount of the precipitation forecast error seen in the control forecast (Fig. 9b) in northwest Queensland. Further experiments nudging Western Australia and the central and western Indian Ocean (not shown), a region where the UM exhibits precipitation biases, have similar skill. This suggests that any forecast errors in these remote source regions are not the primary source of the errors observed in the ensemble forecasts, such that errors are likely either sourced in the immediate vicinity of the low itself or the wider region around northern Australia.

One local model error we have already identified is the land–sea precipitation bias seen in Fig. 8, which is related to the behavior of the convection scheme. If this bias has a major impact on forecast skill, it may be reduced or eliminated by running simulations where the convection scheme is not used. In Fig. 10, we show a series of ensemble forecasts using a nested model suite. In these simulations, a one-way 4.4 km regional

FIG. 10. The 7-day precipitation accumulations (mm) as in Fig. 9, showing (a) AWAP observations and ensemble forecasts from a series of (b),(f),(g) N216 atmosphere-only global ensemble simulations and (c),(e),(g) 4.4-km nested regional model simulations initialized on (b),(c) 25, (d),(e) 30, and (f),(g) 31 Jan.

model (Bush et al. 2020, using a $20^{\circ} \times 20^{\circ}$ domain centered on 20°S , 140°E) is driven at the boundaries by the global N216 atmosphere only model ensemble. Figure 10 shows ensemble-mean nested suite forecasts starting from the 25, 30, and 31 January, with the precipitation from the driving N216 global model for comparison. In the ensembles starting on 25 January (Figs. 10b,c), while there is a clear reduction in the precipitation totals on the western coast of Cape York and a small westward (inland) shift of precipitation on the coast around Townsville, the primary forecast error (the failure to predict the inland accumulations around Mount Isa) remains. This implies that in these forecasts, the error in the precipitation accumulations in the Peak region is either sourced outside of the domain of the nest or is not associated with the coastal precipitation biases caused by the convection scheme, since it is

switched off in these simulations. For example, it may be caused by other processes such as land surface or atmosphere–ocean interactions (as highlighted by the superior performance of the coupled forecasts). A series of experiments where nudging is applied to the model only over the region around northern Queensland (Supplementary Fig. 6) shows similar skill to the globally nudged ensemble (Fig. 9c). These locally nudged experiments would not correct residual errors from outside the nudging domain, such as erroneous moisture advection into the domain from remote circulation errors, but would correct any errors from local parameterizations where the large-scale flow around the storm is well resolved. We therefore conclude that the forecast errors are principally of local origin, either from systematic model biases or from small scale errors growing with time.

To further investigate the source of the forecast error, a series of N216 atmosphere only ensemble simulations are performed, with the forecasts starting daily from 20 to 31 January. The lead time to the start of most intense precipitation (31 January) therefore ranges from zero to 11 days. These forecasts all suffer from the same process biases we have identified above, but provide a tool to identify whether there is significant temporal sensitivity to the evolution of the error. They are shown in Fig. 11. It is apparent that the ensemble forecasts have skill in suggesting an extreme event may occur in the wider North Queensland region in excess of 10 days prior to the date of interest (31 January). For the smaller Peak and Gulf regions, there is little change in forecast skill between the 20 and 30 January, with the confidence intervals (5%–95%, calculated using the *t*-distribution) on the ensemble-mean precipitation accumulations overlapping for all dates. On 31 January, there is a clear uplift in skill, as the ensemble-mean precipitation accumulations for 31 January–6 February more than double from the prior forecasts. This is associated with the forecasts more successfully placing the low south of the Gulf of Carpentaria from 1 to 5 February (see Supplementary Fig. 7), as seen in the observations in Fig. 3, and as was more successfully forecast by the coupled model than atmosphere only simulations (Fig. 7). This indicates a dynamical error evolved in the forecasts prior to 0000 UTC 31 January, which had downstream implications for the evolution and location of the low. We note that this sudden uplift in event probability would be typical of a “sneak” event described by McLay (2011), which, when used in decision-making, are a typically higher cost forecast failures than a “phantom,” where a consistently forecast high-impact event disappears from the forecasts near to the anticipated event occurrence date.

Returning to the nested suites, Figs. 10d–g, show precipitation accumulations from forecasts starting on 30 and 31 January. The nested suites do not show any of the excessive precipitation on the west of Cape York but do not improve the forecast accumulations from the N216 driving model to the south of the Gulf of Carpentaria, where the low stalled and which caused the substantial flooding in northwest Queensland. This implies that, while there is evidently an error in the precipitation accumulations around the coast from the convection scheme, it may not be the source error in terms of the location of the low pressure. This forecast error (the position of the low) which is persistent in atmosphere only ensembles initialized prior to 31 January may be associated with model error, such as atmosphere–surface interactions, or the initial conditions, for example, a lack of key observations to constrain the analysis in an area of particular sensitivity or the impact of assimilating certain observations on the analysis state [see Semple et al. (2012) for an investigation of such sensitivity]. We note that in the forecasts starting on 31 January (Figs. 10f–g), the global model more successfully forecasts the highest precipitation accumulations around Mt. Isa. The dynamical evolution of the forecasts shows these differences are due to the precipitation tailing off in this region on the 4 and 5 February in the 4.4 km model and persisting in the global model (see Supplementary Fig. 8). While the circulation in the global and regional models are comparable throughout

FIG. 11. The 7-day precipitation accumulations (mm) for the three regions as in Fig. 5, but showing ensemble forecasts for a series of N216 atmosphere-only ensemble simulations initialized on the days specified: from 20 Jan in column (d) to 31 Jan in column (o). Columns (a) and (b) are box-and-whisker plots showing the climatological distribution of 7-day DJF precipitation totals in AWAP observations and MOGREPS ensemble forecasts, respectively. Column (c) is the observed AWAP accumulation from 31 Jan to 6 Feb.

the forecast period (not shown), more precipitation falls on the band of mountains on the east in the regional model (centered on 20°S, 147°E), which prevents moisture penetrating farther westward (not shown). The higher-resolution orography in the nested suite may play a role in this difference.

While a forensic analysis of the residual error source(s) would be possible, the purpose of this paper was to highlight the suite of tools that could be applied to forecast the risk of an

extreme event of this nature and a comparative analysis of the benefits those approaches. Further work will expand this initial case study into a wider testbed across multiple case studies to understand how systematic these differences are.

4. Summary

In this study, we have applied a suite of forecast tools using the Met Office's numerical weather prediction system to assess the predictability of an extreme precipitation event in northern Queensland. The precipitation accumulations associated with the event were unprecedented in many locations. The associated flooding caused considerable losses to agricultural communities in the region. The information available to those stakeholders at the time of the event did not allow for any mitigation strategies to be effected.

We have shown that by applying an ensemble approach to forecasting the event, we have the capacity to provide useful and detailed information on the likelihood of extreme rainfall at a multiweek time scale for this event. We note that the experimental framework we have employed here is not currently operational, but was based on the operational model suites being run (or tested) at the Met Office at the time of the event in early 2019. We have further applied a suite of tools to evaluate error sources in the ensemble forecasts to identify where additional skill might be gained over the lower-resolution (N216) atmosphere only ensemble. Ocean-atmosphere coupling played a clear role in improving forecasts of the trajectory of the low pressure system and the behavior of the convection scheme introduced biases in the distribution of precipitation across northern Queensland, particularly in the atmosphere only simulations. One null result of interest is that neither increasing the resolution of the atmosphere nor improving the physics in the atmosphere model improve the forecast skill in this case study. We also note the nested regional model did not outperform the global model for later start dates (when the global model was more accurate), though these differences only manifested themselves late in the forecasts and may not represent a systematic underperformance by the nested model. Understanding the details of these differences would be an important next step to establish the utility of running the computationally expensive nested suite for this region.

Nudged experiments indicate that good spatial, temporal, and amplitudinal accumulations of precipitation can be generated by the parameterization schemes if the large-scale flow is well represented. The errors in the large-scale flow may well be generated by the growth of small errors from the model physics interacting and growing rapidly—for this event, the failure of the forecast skill to improve when remote nudging was performed indicates that the error growth is not remotely sourced. Such a conclusion reconciles with the limits of predictability found for this event by Cowan et al. (2019), where an ability to forecast the broad scale atmospheric conditions around Northern Australia (including the state of ENSO, the MJO and the SAM) did not correlate with similar skill in forecasting the location and intensity of the observed intense precipitation in northwest Queensland. It further reconciles

with the step change in forecast skill found in the lower-resolution (N216) ensemble from the 31 January onward—if the error was remotely sourced, it might be expected that any such improvements in skill would occur with a lead time equivalent to the time it would take a remotely sourced error to propagate into the region.

To operationalize a multiweek ensemble approach as used here, which would have particular utility in identifying the likelihood of extreme events beyond the 7-day forecast period that is often used operationally, forecast centers would need to rapidly process a substantial volume of data from both observational and model climatologies (with bias correction) in real time in order to extract a relatively small signal in the tails of distributions. For this event, the BOM's monthly rainfall outlook for February 2019 was issued on 31 January, but only included forecasts started between 18 and 26 January, where there was lower skill (see Cowan et al. 2019), demonstrative of the difficult balance forecast centers need to strike between timeliness and provision of detailed information. We note that in an operational context, including bias correction would improve the long-term performance of the ensemble in predicting the magnitude of extreme events, though for individual extreme events may not always improve the accuracy of those forecasts (e.g., Pantillon et al. 2018). Here, we perform the analysis for an identified location, but for real time event identification it would need to cover a wider forecast domain and produce statistics across that domain, and this presents a considerable challenge. In this context, we note that in developing an operational forecast system it would be verified over many cases. This study represents a first look at the potential benefits of a multiweek ensemble approach. To properly assess the benefits of this approach it would need to be verified probabilistically. Here, reference to the MOGREPS ensemble is used as an available and practical approach to provide a baseline to assess forecast skill. We note, however, that the distribution of a hypothetical week 2 MOGREPS forecast climatology may differ to the week 1 distribution evaluated in this study.

Further work will expand this case study into a wider testbed in order to separate systematic effects from those that may be unique to this case study. Such an analysis will assist in identifying priorities for model development while informing optimal approaches to real-time forecasting of extreme event risk at multiweek time scales.

Acknowledgments. MH, SL, and TC were funded by Meat and Livestock Australia (MLA), the Queensland Government through the Drought and Climate Adaptation Program, and University of Southern Queensland through the Northern Australian Climate Program (NACP). The NACP is coordinated by Roger Stone. DC was supported by the Met Office Hadley Centre Climate Programme funded by BEIS and Defra.

Data availability statement. All observational and reanalysis data are publicly available. Data from the simulations are available from MKH upon reasonable request.

REFERENCES

- Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, <https://doi.org/10.1038/nature14956>.
- Belcher, S. E., H. Hewitt, A. Beljaars, E. Brun, B. Fox-Kemper, J.-F. Lemieux, G. Smith, and S. Valcke, 2015: Ocean-waves-sea ice-atmosphere interactions. *Seamless Prediction of the Earth System: From Minutes to Months*, G. Brunet, S. Jones, and P. Ruti, Eds., WMO, 155–169.
- Birch, C. E., M. J. Roberts, L. Garcia-Carreras, D. Ackerley, M. J. Reeder, A. P. Lock, and R. Schiemann, 2015: Sea-breeze dynamics and convection initiation: The influence of convective parameterization in weather and climate model biases. *J. Climate*, **28**, 8093–8108, <https://doi.org/10.1175/JCLI-D-14-00850.1>.
- Bishop, C. H., B. J. Etherton, and S. J. Majumdar, 2001: Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. *Mon. Wea. Rev.*, **129**, 420–436, [https://doi.org/10.1175/1520-0493\(2001\)129<0420:ASWTET>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<0420:ASWTET>2.0.CO;2).
- Blockley, E., and Coauthors, 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, <https://doi.org/10.5194/gmd-7-2613-2014>.
- Bowler, N. E., A. Arribas, K. R. Mylne, K. B. Robertson, and S. E. Beare, 2008: The MOGREPS short-range ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **134**, 703–722, <https://doi.org/10.1002/qj.234>.
- Brassington, G., and Coauthors, 2015: Progress and challenges in short- to medium-range coupled prediction. *J. Operat. Oceanogr.*, **8**, s239–s258, <https://doi.org/10.1080/1755876X.2015.1049875>.
- Buizza, R., P. Houtekamer, G. Pellerin, Z. Toth, Y. Zhu, and M. Wei, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.*, **133**, 1076–1097, <https://doi.org/10.1175/MWR2905.1>.
- Bureau of Meteorology, 2019a: Special Climate Statement 68—Widespread heatwaves during December 2018 and January 2019. Bureau of Meteorology, 70 pp., <http://www.bom.gov.au/climate/current/statements/scs68.pdf>.
- , 2019b: Special Climate Statement 69—An extended period of heavy rainfall and flooding in tropical Queensland. Bureau of Meteorology, 47 pp., <http://www.bom.gov.au/climate/current/statements/scs69.pdf>.
- Bush, M., and Coauthors, 2020: The first Met Office Unified Model-JULES Regional Atmosphere and Land configuration, RAL1. *Geosci. Model Dev.*, **13**, 1999–2029, <https://doi.org/10.5194/gmd-13-1999-2020>.
- Clark, P., N. Roberts, H. Lean, S. P. Ballard, and C. Charlton-Perez, 2016: Convection-permitting models: A step-change in rainfall forecasting. *Meteor. Appl.*, **23**, 165–181, <https://doi.org/10.1002/met.1538>.
- Cowan, T., and Coauthors, 2019: Forecasting the extreme rainfall, low temperatures, and strong winds associated with the northern Queensland floods of February 2019. *Wea. Climate Extremes*, **26**, 100232, <https://doi.org/10.1016/j.wace.2019.100232>.
- GEBCO Compilation Group, 2020: GEBCO_2020 Grid. Accessed 20 April 2021, <https://doi.org/10.5285/a29c5465-b138-234d-e053-6c86abc040b9>.
- Hall, T., J. Milson, and C. Hall, 2020: Pasture recovery, land condition and some other observations after the monsoon flooding, chill event in north-west Queensland in Jan-Mar 2019. Tech. Rep., Department of Agriculture and Fisheries, State of Queensland, 45 pp., <http://era.daf.qld.gov.au/id/eprint/7443/>.
- Hewitt, H., D. Copsey, I. Culverwell, C. Harris, R. Hill, A. Keen, A. McLaren, and E. Hunke, 2011: Design and implementation of the infrastructure of HadGEM3: The next-generation Met Office climate modelling system. *Geosci. Model Dev.*, **4**, 223–253, <https://doi.org/10.5194/gmd-4-223-2011>.
- Hou, A. Y., and Coauthors, 2014: The Global Precipitation Measurement Mission. *Bull. Amer. Meteor. Soc.*, **95**, 701–722, <https://doi.org/10.1175/BAMS-D-13-00164.1>.
- Huffman, G. J., R. F. Adler, M. M. Morrissey, D. T. Bolvin, S. Curtis, R. Joyce, B. McGavock, and J. Susskind, 2001: Global precipitation at one-degree daily resolution from multisatellite observations. *J. Hydrometeorol.*, **2**, 36–50, [https://doi.org/10.1175/1525-7541\(2001\)002<0036:GPAODD>2.0.CO;2](https://doi.org/10.1175/1525-7541(2001)002<0036:GPAODD>2.0.CO;2).
- Hunke, E., W. Lipscomb, A. Turner, N. Jeffery, and S. Elliott, 2015: CICE: The Los Alamos sea ice model documentation and software user's manual version 5. Tech. Rep. LA-CC-06-012, Los Alamos National Laboratory, Los Alamos, NM, 116 pp.
- Johnson, B., J. Haywood, and M. Hawcroft, 2019: Are changes in atmospheric circulation important for black carbon aerosol impacts on clouds, precipitation, and radiation? *J. Geophys. Res. Atmos.*, **124**, 7930–7950, <https://doi.org/10.1029/2019JD030568>.
- Jones, D. A., W. Wang, and R. Fawcett, 2009: High-quality spatial climate data-sets for Australia. *Aust. Meteor. Oceanogr. J.*, **58**, 233–248, <https://doi.org/10.22499/2.5804.003>.
- Kendon, E. J., N. M. Roberts, C. A. Senior, and M. J. Roberts, 2012: Realism of rainfall in a very high-resolution regional climate model. *J. Climate*, **25**, 5791–5806, <https://doi.org/10.1175/JCLI-D-11-00562.1>.
- MacLachlan, C., and Coauthors, 2015: Global Seasonal forecast system version 5 (GloSea5): A high-resolution seasonal forecast system. *Quart. J. Roy. Meteor. Soc.*, **141**, 1072–1084, <https://doi.org/10.1002/qj.2396>.
- Madeo, G., and the NEMO team, 2016: NEMO Ocean Engine. Note du Pole de modelisation de l'Institut Pierre-Simon Laplace, Rep. 27, 386 pp.
- McLay, J., 2011: Diagnosing the relative impact of sneaks, phantoms, and volatility in sequences of lagged ensemble probability forecasts with a simple dynamic decision model. *Mon. Wea. Rev.*, **139**, 387–402, <https://doi.org/10.1175/2010MWR3449.1>.
- Pantillon, F., S. Lerch, P. Knippertz, and U. Corsmeier, 2018: Forecasting wind gusts in winter storms using a calibrated convection-permitting ensemble. *Quart. J. Roy. Meteor. Soc.*, **144**, 1864–1881, <https://doi.org/10.1002/qj.3380>.
- Rauniyar, S., A. Protat, and H. Kanamori, 2017: Uncertainties in TRMM-era multisatellite-based tropical rainfall estimates over the Maritime Continent. *Earth Space Sci.*, **4**, 275–302, <https://doi.org/10.1002/2017EA000279>.
- Rawlins, F., S. Ballard, K. Bovis, A. Clayton, D. Li, G. Inverarity, A. Lorenc, and T. Payne, 2007: The Met Office global four-dimensional variational data assimilation scheme. *Quart. J. Roy. Meteor. Soc.*, **133**, 347–362, <https://doi.org/10.1002/qj.32>.
- Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, <https://doi.org/10.1175/2007MWR2123.1>.
- Rodríguez, J. M., and S. F. Milton, 2019: East Asian summer atmospheric moisture transport and its response to interannual variability of the West Pacific subtropical high: An evaluation of the Met Office Unified Model. *Atmosphere*, **10**, 457, <https://doi.org/10.3390/atmos10080457>.
- Schwartz, C. S., and R. A. Sobash, 2017: Generating probabilistic forecasts from convection-allowing ensembles using

- neighborhood approaches: A review and recommendations. *Mon. Wea. Rev.*, **145**, 3397–3418, <https://doi.org/10.1175/MWR-D-16-0400.1>.
- Semple, A., M. Thurlow, and S. Milton, 2012: Experimental determination of forecast sensitivity and the degradation of forecasts through the assimilation of good quality data. *Mon. Wea. Rev.*, **140**, 2253–2269, <https://doi.org/10.1175/MWR-D-11-00273.1>.
- Smith, G. C., and Coauthors, 2018: Impact of coupling with an ice–ocean model on global medium-range NWP forecast skill. *Mon. Wea. Rev.*, **146**, 1157–1180, <https://doi.org/10.1175/MWR-D-17-0157.1>.
- Sorooshian, S., K.-L. Hsu, X. Gao, H. V. Gupta, B. Imam, and D. Braithwaite, 2000: Evaluation of PERSIANN system satellite-based estimates of tropical rainfall. *Bull. Amer. Meteor. Soc.*, **81**, 2035–2046, [https://doi.org/10.1175/1520-0477\(2000\)081<2035:EOPSSE>2.3.CO;2](https://doi.org/10.1175/1520-0477(2000)081<2035:EOPSSE>2.3.CO;2).
- Valcke, S., T. Craig, and L. Coquart, 2013: OASIS3-MCT user guide, OASIS3-MCT 2.0. CERFACS/CNRS SUC URA, 50 pp.
- Vellinga, M., D. Copsey, T. Graham, S. Milton, and T. Johns, 2020: Evaluating benefits of two-way ocean–atmosphere coupling for global NWP forecasts. *Wea. Forecasting*, **35**, 2127–2144, <https://doi.org/10.1175/WAF-D-20-0035.1>.
- Vitart, F., and Coauthors, 2017: The Subseasonal to Seasonal (S2S) prediction project database. *Bull. Amer. Meteor. Soc.*, **98**, 163–173, <https://doi.org/10.1175/BAMS-D-16-0017.1>.
- Walters, D., and Coauthors, 2017: The Met Office Unified Model global atmosphere 6.0/6.1 and JULES global land 6.0/6.1 configurations. *Geosci. Model Dev.*, **10**, 1487–1520, <https://doi.org/10.5194/gmd-10-1487-2017>.
- Williams, K., and Coauthors, 2018: The Met Office global coupled model 3.0 and 3.1 (GC3.0 and GC3.1) configurations. *J. Adv. Model. Earth Syst.*, **10**, 357–380, <https://doi.org/10.1002/2017MS001115>.
- Xiang, B., M. Zhao, X. Jiang, S.-J. Lin, T. Li, X. Fu, and G. Vecchi, 2015: The 3–4-week MJO prediction skill in a GFDL coupled model. *J. Climate*, **28**, 5351–5364, <https://doi.org/10.1175/JCLI-D-15-0102.1>.