Identifying Periods of Forecast Model Confidence for Improved Subseasonal Prediction of Precipitation

DOUG RICHARDSON,^a AMANDA S. BLACK,^a DIDIER P. MONSELESAN,^a THOMAS S. MOORE II,^a JAMES S. RISBEY,^a ANDREW SCHEPEN,^b DOUGAL T. SQUIRE,^a AND CARLY R. TOZER^a

^aCSIRO Oceans and Atmosphere, Hobart, Tasmania, Australia

^bCSIRO Land and Water, Brisbane, Queensland, Australia

(Manuscript received 24 February 2020, in final form 18 November 2020)

ABSTRACT: Subseasonal forecast skill is not homogeneous in time, and prior assessment of the likely forecast skill would be valuable for end-users. We propose a method for identifying periods of high forecast confidence using atmospheric circulation patterns, with an application to southern Australia precipitation. In particular, we use archetypal analysis to derive six patterns, called archetypes, of daily 500-hPa geopotential height (Z_{500}) fields over Australia. We assign Z_{500} reanalysis fields to the closest-matching archetype and subsequently link the archetypes to precipitation for three key regions in the Australian agriculture and energy sectors: the Murray Basin, southwest Western Australia, and western Tasmania. Using a 20-yr hindcast dataset from the European Centre for Medium-Range Weather Forecasts subseasonal-to-seasonal prediction system, we identify periods of high confidence as when hindcast Z_{500} fields closely match an archetype according to a distance criterion. We compare the precipitation hindcast accuracy during these confident periods compared to normal. Considering all archetypes, we show that there is greater skill during confident periods for lead times of less than 10 days in the Murray Basin and western Tasmania, and for greater than 6 days in southwest Western Australia, although these conclusions are subject to substantial uncertainty. By breaking down the skill results for each archetype individually, we highlight how skill tends to be greater than normal for those archetypes associated with drier-than-average conditions.

KEYWORDS: Australia; Synoptic climatology; Forecast verification/skill; Probabilistic Quantitative Precipitation Forecasting (PQPF); Seasonal forecasting; Clustering

1. Introduction

Subseasonal prediction skill of extratropical climate variables such as precipitation is typically low overall (de Andrade et al. 2019). Skill, however, is not homogeneous in time, and it would be useful for end-users to assess the likely skill of a forecast a priori. The concept of “forecasting forecast skill” originated several decades ago, focusing on short-range (up to 10 days) forecasts of atmospheric variables (Kalnay and Dalcher 1987; Dalcher et al. 1988; Palmer and Tibaldi 1988; Molteni and Palmer 1991; Wobus and Kalnay 1995); see section 2 of Kalnay (2019) for a historical overview of the topic. A common approach taken by those studies was to exploit the postulated relationship between forecast skill and spread, whereby the average correlation between ensemble members of a forecast was used as a predictor of forecast skill. Another approach is assessing the “flow-dependent predictability,” whereby forecasts are stratified by the broadscale initial state of the atmosphere. This framework allows for a comparison of

forecast skill arising from forecasts initialized during different atmospheric states (e.g., the state of some large-scale climate mode). The subseasonal-to-seasonal (S2S) prediction community (Vitart et al. 2017) has contributed research in this regard, often with reference to forecast “windows of opportunity” (e.g., Rodwell and Doblas-Reyes 2006; Hoskins 2013; Mariotti et al. 2020), during which there is a tendency for greater-than-normal skill. Such periods are generally characterized by climate states at the forecast initialization time that favor a more predictable evolution over the lead time, such as the state of El Niño–Southern Oscillation (ENSO; Qin and Robinson 1995; Branković and Palmer 2000; Shukla et al. 2000; Goddard and Dilley 2005; Frías et al. 2010; Kim et al. 2012; Manzananas et al. 2014; Miller and Wang 2019), phases of the Madden–Julian oscillation (MJO; Jones et al. 2004; Neena et al. 2014; Jones et al. 2015; Ferranti et al. 2018), or particular patterns of atmospheric circulation (Frame et al. 2013; Ferranti et al. 2015, 2018; Vigaud et al. 2018).

By adopting an atmospheric circulation pattern (CP) approach, we propose a new methodology for identifying periods of forecast confidence. CP classifications are datasets containing a number of distinct states, the CPs, that represent the circulation over a specified domain and time scale (Huth et al. 2008). These CPs are typically derived by applying some clustering algorithm to a dataset of atmospheric fields, often daily mean sea level pressure (MSLP) or 500-hPa geopotential height (Z_{500}). They can be a useful tool for reducing the dimensionality of an analysis, as a continuous time series of, for example, Z_{500} reanalysis fields reduces to a discrete time series of CP occurrences. Predominantly used in the field of synoptic

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/JHM-D-20-0054.s1>.

Corresponding author: Doug Richardson, doug.richardson@csiro.au

DOI: 10.1175/JHM-D-20-0054.1

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

climatology (Huth et al. 2008), CP classifications have recently been utilized in forecasting applications, such as the prediction of extreme events (Muñoz et al. 2016; Ferranti et al. 2018; Lavaysse et al. 2018; Neal et al. 2018; Richardson et al. 2020a,b), model bias correction (Vuillaume and Herath 2017) and forecast skill evaluation (Frame et al. 2013; Ferranti et al. 2015).

The term “circulation pattern” is seemingly interchangeable with any combination of the words “circulation” or “weather” and “pattern” or “type.” These classifications can consist of tens of CPs, each of which tend to persist for only a few consecutive days (e.g., Lamb 1972; Bogardi et al. 1993; Casado et al. 2010; Neal et al. 2016). On the other hand, “weather regime” classifications tend to contain fewer patterns, and are intended to capture longer time scale, quasi-stationary states rather than daily atmospheric circulation variability (e.g., Michelangeli et al. 1995; Stephenson et al. 2004).

CPs (or similar) have also been used in a subseasonal forecast “windows of opportunity” context. Frame et al. (2013) characterized the North Atlantic eddy-driven jet using the leading two empirical orthogonal functions (EOFs), and showed that the winter predictability of the jet was lower when associated with the EOF space corresponding to weak or split jets. A two-dimensional EOF phase space was also used by Ferranti et al. (2018), who studied the relationship between midtropospheric flow and severe winter cold temperature anomalies over Europe. The authors found that predictability is greater during a negative phase of the North Atlantic Oscillation (NAO), and that hindcasts initialized during a MJO event (any phase) exhibit higher skill in predicting cold extremes. Similarly, Ferranti et al. (2015) showed that winter hindcasts of Z_{500} over the Euro-Atlantic domain were more skillful when initialized during a negative NAO-like regime. In this case the authors considered four weather regimes (positive NAO, negative NAO, Scandinavian blocking, and an Atlantic Ocean ridge) generated by applying a principal components analysis (PCA) followed by K -means clustering to Z_{500} fields from the European Centre for Medium-Range Weather Forecasts (ECMWF) operational reanalysis.

In this study we take a different approach. Rather than assessing skill stratified by the CP at initialization time, we hindcast CPs explicitly, creating subsamples of predictions for each CP from which precipitation hindcast skill can be assessed. This is a similar framework to that used by Nigro et al. (2011), who assessed the 3-hourly hindcast skill (up to 72 h) of a number of surface variables from the Antarctic Mesoscale Prediction System, with self-organizing maps (Kohonen 2001) used to generate the CPs.

We extend this analysis design by further subsampling based on how well matched the CPs are to the underlying data. This allows us to calculate the precipitation hindcast skill for periods when the model has high confidence of the hindcast CP relative to periods when the model is less confident about the hindcast CP. We also show that the additional subsampling step modifies the observed conditional distributions of precipitation given the CPs, achieving a greater distinction between the precipitation associated with each CP. The CPs are generated using archetypal analysis (AA; Cutler and Breiman 1994),

which is relatively new to the climate science community, and is theoretically better suited to dealing with extremes than traditional CP-generating methods such as PCA or K -means clustering (Steinschneider and Lall 2015; Hannachi and Trendafilov 2017).

We frame this article around a particular application: sub-seasonal prediction, using the ECMWF extended-range ensemble prediction system, of daily precipitation for three regions in southern Australia. Previous studies have shown that this system has virtually no hindcast skill for southern Australia precipitation beyond a 1-week lead time (Li and Robertson 2015; de Andrade et al. 2019). This region is therefore a useful testbed for exploring the utility of our proposed method.

This study is organized as follows. In section 2 we detail the datasets used and the methods for generating the archetypes, deriving the archetype reanalysis time series and hindcast dataset, relating the archetypes to precipitation, defining periods of high forecast confidence and assessing the precipitation hindcast skill. Section 3 describes the results, and section 4 provides a discussion and concluding remarks.

2. Data and methods

a. Generating archetypes

Conventional methods for generating a CP classification, such as PCA or K -means clustering, focus on the majority of the probability distribution of the data they are applied to, for example, by estimating the data centroids. AA instead estimates the convex hull, or “corners” of the data, and therefore yields features, called archetypes, that represent the extremes (Cutler and Breiman 1994; Mørup and Hansen 2012). An advantage of AA is that, unlike for K -means or PCA, the original data can be approximated by a convex combination of the archetypes. Despite the prevalence of research into climate extremes, AA was only recently introduced to this community. Steinschneider and Lall (2015) generated archetypes based on two daily precipitation variables over the eastern United States and related them to atmospheric circulation anomalies and SST teleconnections, while Hannachi and Trendafilov (2017) presented a new AA algorithm based on nonlinear optimization on manifolds (Boumal et al. 2014) with an application to ENSO and the Asian summer monsoon.

We generate a set of archetypes from daily Z_{500} anomaly fields over Australia and part of the Southern Ocean (10° – 60° S, 100° – 170° E) from the Japanese 55-Year Reanalysis (JRA55; Japan Meteorological Agency 2013) between 1958 and 2016 at a resolution of 1.5° latitude and longitude (regridged from a 1.25° resolution). The anomalies are obtained by subtracting the daily JRA55 1958–2016 annual-cycle climatology from the data.

Before implementing AA, we carry out a PCA on the anomalies, retaining 13 principal components that together explain 95% of the total variance. The PCA step is purely to reduce the dimensionality of the data to improve computational performance. We implement AA using the MATLAB routine presented by Mørup and Hansen (2012). This routine is applied to the scaled principal components, denoted by $\mathbf{X}$. For a predefined number of archetypes k , AA then attempts to

find two matrices, $\mathbf{C}$ and $\mathbf{S}$, that minimize the squared Frobenius norm given by

$$\|\mathbf{X} - \mathbf{XCS}\|_F^2 = \text{trace}[(\mathbf{X} - \mathbf{XCS})^T(\mathbf{X} - \mathbf{XCS})]. \quad (1)$$

The matrices $\mathbf{C}$ and $\mathbf{S}$ are stochastic matrices that can be readily interpreted as probabilities, i.e., $\sum_{i=1}^I C_{ij} = 1$ and $\sum_{i=1}^I S_{ij} = 1$, $\forall j$, with $\mathbf{C} \geq 0$ and $\mathbf{S} \geq 0$. These constraints enforce the matrix $\mathbf{XC}$ to be a weighted average of the data, and for each data point to be approximated by a weighted average of $\mathbf{XC}$ (Mørup and Hansen 2012). The last step is to reconstruct $\mathbf{XC}$ into the original units and dimensions by multiplication with the EOFs, $\mathbf{E}$. We thus obtain our final archetype definitions, $\mathbf{EXC}$ (i.e., the CPs, referred to from here as the archetypes).

The implications of the constraints on $\mathbf{C}$ and $\mathbf{S}$ apply to the archetypes as well as the PCs on which AA was performed. The archetypes, then, are weighted averages of the JRA55 Z_{500} anomaly fields. In common with Hannachi and Trendafilov (2017), we find that the number of daily anomaly fields that contribute to the archetype definitions is small with respect to the length of the input data record (not shown). That is to say, a small number of daily Z_{500} anomaly fields are used to define the archetypes. The implication of the constraints on $\mathbf{S}$ mean that each daily Z_{500} anomaly field in JRA55 can be approximated as a weighted combination of the archetypes. As is standard with many optimization problems, AA should be run a number of times (in our case 1000) with different initial conditions. The run with the best-scoring optimization criterion is then selected for further analysis.

Despite the majority of modern CP classifications being labeled as “objective,” there are typically several subjective decisions that must be made, such as the domain size, resolution and the desired number of CPs (Huth et al. 2008). This last decision is perhaps the most fraught to justify, and the various suggested indicators of an optimal number of CPs does not always yield a clear answer, most likely because there is no predetermined number of CPs in circulation datasets (e.g., Stephenson et al. 2004; Christiansen 2007; Philipp et al. 2007; Fereday et al. 2008). Indeed, many of these indicators are generic methods for identifying the optimum number of clusters from some clustering algorithm (Milligan and Cooper 1985; Michelangeli et al. 1995; Gerstengarbe and Werner 1997; Hannachi et al. 2017). This focus on how well the clusters represent the data is understandable for studies primarily interested in achieving the most realistic classification. Other studies have taken a more application-focused approach to selecting the number of CPs: the choice is generally considered a balance between choosing too few CPs, which may result in large intra-CP variability (e.g., in how well the CPs distinguish between precipitation patterns), and too many CPs, for which there may be a number of CPs that do not appear distinct from each other and small sample sizes of each CP (e.g., Nigro et al. 2011; Neal et al. 2016). Having experimented with several choices of the number of archetypes, we find six archetypes to be optimal (see section 4).

b. Deriving archetype time series and hindcasts

Having generated the archetype definitions, we derive an observed archetype time series and a hindcast dataset. The

approach in constructing these two datasets is to assign daily Z_{500} anomaly fields to the closest-matching archetype by minimizing the sum of squared differences (SSD) between the field and each archetype definition. For the observed time series, the fields are from the JRA55 reanalysis, on the same grid and over the same period as for the archetype generation. For the hindcast dataset, we use the ECMWF ensemble prediction system from the S2S prediction project (ECMWF-S2S; Vitart et al. 2017). This is a coupled atmosphere–ocean–sea ice model with a lead time of 46 days. The horizontal atmospheric resolution is ~ 16 km up to a lead time of 15 days and ~ 32 km beyond this. It is run twice weekly (Mondays and Thursdays) at 0000 UTC with 11 ensemble members available for hindcasts. We extract hindcast data for Z_{500} , over the same domain and at the same grid resolution as used for the archetype definitions and time series, for initialization dates between 2 January 1997 and 28 December 2016, inclusive, giving a total of 2080 initialization dates. The data are converted to anomalies by subtracting the Z_{500} ensemble mean, dependent on both initialization date and lead time. It is these fields that are assigned to the closest archetype.

c. Subsampling archetypes

In typical CP studies, a CP is assigned to every field in a dataset by determining the CP in the classification that is the closest match to the target field according to some set of criteria (e.g., Philipp et al. 2007; Casado et al. 2010; Nigro et al. 2011; Neal et al. 2016; Richardson et al. 2019). Therefore, even in cases when none of the CPs are particularly similar to a given field, a CP will be assigned regardless. For one of the earliest CP classifications, Lamb (1972) predefined MSLP CPs based on the dominant flow direction and/or cyclonicity (e.g., northerly, southwesterly, cyclonic, anticyclonic) and subjectively classified the daily MSLP fields. On those relatively rare occasions when the field did not appear to fit any of the CPs, the day was categorized as “unclassified.” When Jenkinson and Collison (1977) designed an objective method for assigning these CPs to fields, unclassified days were determined to be when estimates for the flow and shear vorticity were less than some threshold.

We take a different approach when assessing the fit of the archetypes to the underlying Z_{500} anomaly fields. As described earlier, the archetype assignment procedure is to minimize the SSD between the field and each archetype. Therefore, we have a probability distribution of the minimum SSD values at each time step for both JRA55 and the ECMWF-S2S hindcast dataset (the latter also has SSD values for each ensemble member). Then, by selecting percentile thresholds of these distributions, we subsample from the two datasets, retaining only those days when the SSD is less than the threshold i.e., when the archetype is a good fit to the field. We set the threshold as the 10th percentile of these distributions, determined separately for JRA55 and the hindcast data. Similarly, we subsample the upper 10% of the SSD distribution to obtain cases where the archetypes are the least well matched to the fields.

We will discuss results in terms of “confidence.” Confident periods, or periods of high confidence, refer to when the data satisfy the lower-tail SSD threshold and hence are “confident”

FIG. 1. Archetype definitions and precipitation distributions. The regions are shown in the top row. The JRA55 500-hPa geopotential height (Z_{500}) climatology (m) is shown in the top-right panel. The six archetypes, generated using data from 1958 to 2016, are shown in the right-hand column between 1958 and 2016. Blue/red shading shows the Z_{500} anomalies (m) and the contours are the full fields, spaced every 100 m. The archetype full fields are generated by adding the archetype anomalies to the Z_{500} climatology. The sample size (days) and percentage occurrence of each archetype are shown to the right, for all days (A) and for confident days (C). (a)–(r) Each archetype's associated 1958–2016 precipitation distribution is shown. The gray dashed line is the climatological CDF, the blue line is the CDF for all days corresponding to a given archetype, and the orange line is the CDF for confident days.

about the assigned archetype. Normal periods, on the other hand, are when the lower-tail SSD threshold is not met, and the data do not look sufficiently like any archetype to be considered confident. Finally, periods of low confidence are when the data lie above the upper-tail SSD threshold, indicating they are not at all well matched to any archetype.

d. Relating archetypes to precipitation

We use daily total precipitation for the period 1958 through 2017 from the Australian Water Availability Project (AWAP) dataset (Jones et al. 2009; Raupach et al. 2009), which applies an interpolation scheme to in situ observations to derive a gridded product with a resolution of 0.05° latitude and longitude. In this analysis we focus on three regions in southern Australia: the Murray Basin, southwest Western Australia, and western Tasmania (Fig. 1). These regions correspond to three of the 15 natural resource management “sub-clusters” (CSIRO and Bureau of Meteorology 2015), which are designed to delineate different climates and can be considered spatially homogeneous in precipitation. These three regions are important for Australia’s agriculture, water resources and hydropower sectors. We calculate a single time series for each of these regions by applying a mask to the AWAP data, and computing the mean precipitation from the cells whose centers are located within the regions’ boundaries. Conditional distributions of observed regional precipitation given each archetype are obtained by stratifying precipitation values by archetype from the JRA55-derived time series. This allows an assessment of the likely daily precipitation given the archetype.

e. Precipitation hindcasts and skill assessment

We obtain hindcast daily precipitation totals from ECMWF-S2S for Australia with a resolution of 0.5° latitude and longitude for the same initialization dates as the Z_{500} hindcasts (section 2b). The data are converted to hindcasts of regional precipitation for the three regions using the same process as described for the observed precipitation in section 2d. We calibrate these hindcasts using quantile mapping, which matches the cumulative distribution function (CDF) of the raw forecasts to the CDF of the observations, in our case the regional precipitation data derived from AWAP. Quantile mapping can be achieved parametrically by fitting a distribution to the data (e.g., Piani et al. 2010; Zhao et al. 2017) or nonparametrically using, for example, empirical CDFs (e.g., Gudmundsson et al. 2012; Bennett et al. 2014; Ratri et al. 2019). We take the latter approach, applying the procedure to each ensemble member separately, within a leave-one-year-out cross validation framework. The empirical CDFs are constructed separately for each month to account for seasonality. In cases where a predicted value exceeds the maximum of the observations, we use linear extrapolation to obtain our calibrated value (Gudmundsson et al. 2012). If necessary this extrapolated value is clipped to the observed maximum multiplied by two standard deviations to prevent grossly inflated calibrated values.

Our primary aim is to compare the hindcast precipitation skill of ECMWF-S2S during confident periods compared to both normal and low confidence periods. We use the mean squared error to assess skill, which, at a given lead time, is defined as

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (f_i - o_i)^2, \quad (2)$$

where N is the hindcast sample size, the f_i are the hindcasts and the o_i are the observations. We can compute the MSE for both confident and normal/low confidence periods, denoted $\text{MSE}^{(c)}$ and $\text{MSE}^{(u)}$, respectively. We calculate the ratio of these two quantities as

$$r = \frac{\text{MSE}^{(c)}}{\text{MSE}^{(u)}}. \quad (3)$$

Therefore, when $r < 1$ the model is more skillful during confident periods compared to either normal or low confidence periods.

When comparing periods of high confidence to normal periods, the sample size of the former is smaller than the sample size of the latter. Therefore we compute $\text{MSE}^{(u)}$ on a random sample of size equal to the sample size of $\text{MSE}^{(c)}$. Furthermore, by repeating the random sampling (i.e., bootstrap resampling with replacement) of $\text{MSE}^{(u)}$ 1000 times we can construct confidence intervals around r . When comparing periods of high and low confidence, the sample sizes of $\text{MSE}^{(c)}$ and $\text{MSE}^{(u)}$ are equal and so no bootstrap resampling is possible.

It is important to highlight that this process is applied separately to each ensemble member i.e., we are calculating skill in a deterministic sense for each member. This is necessary because for a given forecast (i.e., for a particular initialization date and lead time), it is possible (probable, even, at long lead times when the ensemble has diverged) that very few members will simultaneously predict an archetype with high confidence. Calculating probabilistic skill scores is therefore unrealistic due to small sample sizes. We instead account for uncertainty by calculating the 5th, 50th, and 95th percentiles of the bootstrapped (for normal periods) ensemble r at each lead time. The median provides the central estimate for skill, while the 5th and 95th percentiles are used for confidence intervals.

A primary reason we use the MSE is that it can be linearly decomposed to provide a skill assessment for each archetype individually. The squared errors term in Eq. (2) is a set composed of the squared errors associated with each individual archetype $k = 1, 2, \dots, 6$. That is,

$$\sum_{i=1}^N (f_i - o_i)^2 = \sum_{k=1}^6 \sum_{i=1}^{N_k} (f_{k,i} - o_{k,i})^2, \quad (4)$$

where the $f_{k,i}$ and $o_{k,i}$ are the hindcasts and observations, respectively, for a specific archetype k , and N_k is the sample size for archetype k , such that $N = \sum_k N_k$. Equation (2) can then be written as

$$\text{MSE} = \frac{1}{N} \sum_{k=1}^6 \frac{\sum_{i=1}^{N_k} (f_{k,i} - o_{k,i})^2}{N_k}, \quad (5)$$

$$= \frac{1}{N} \sum_{k=1}^6 N_k \text{MSE}_k, \quad (6)$$

and therefore Eq. (3) becomes

$$r = \frac{\sum_{k=1}^6 N_k^{(c)} \text{MSE}_k^{(c)}}{N^{(c)} \text{MSE}^{(u)}} \quad (7)$$

$$= \sum_{k=1}^6 r_k. \quad (8)$$

The superscripts (c) and (u) are again used to associate the relevant variables to confident and normal/low confidence periods. For individual archetypes, the model is more skillful during confident periods when

$$\frac{\text{MSE}_k^{(c)}}{\text{MSE}^{(u)}} < 1, \quad (9)$$

which, using Eqs. (7) and (8), is equivalent to

$$r_k < \frac{N_k^{(c)}}{N^{(c)}}. \quad (10)$$

Note that this decomposition only works if $\text{MSE}^{(u)}$ is independent of k . Strong archetype seasonality would violate this independence, as the random sampling for $\text{MSE}^{(u)}$ would need to account for seasonality, and would therefore differ between archetypes. We find very little seasonality evident in the archetype occurrences (see section 3a).

Comparing precipitation skill during periods of high confidence with skill during normal or low confidence periods represents two kinds of test. The comparison of high and low confidence periods tests how well the method distinguishes precipitation skill based on its ability to identify forecast fields that look like archetypes. When comparing periods of high confidence to normal, however, we are applying a much stricter test by asking how useful the is method compared an archetype-agnostic process.

f. Dry days test

As we will show, the results imply that there is greater skill during confident predictions of archetypes when the archetype is associated with drier-than-average conditions for a particular region. We therefore develop a test to assess whether the gain in skill in such situations is a result of the archetype framework, or due to dry days in the hindcast being more accurate regardless (i.e., independent of confidence in an archetype).

For a given lead time, region and archetype, we take the maximum hindcast precipitation value from the subset of days that coincide with a confident prediction of that archetype, $p_{\max}$, and calculate the MSE from these days. We then take 1000 bootstrap samples from the hindcast precipitation dataset for values equal to or less than $p_{\max}$, and calculate the MSE for each of these samples. This is repeated for each ensemble member, yielding two distributions: one comprising the 11 MSE scores corresponding to confident predictions of the archetype, and the other comprising the 11 000 bootstrap samples.

If the archetype framework is effective in identifying situations of greater-than-normal skill, we would expect the smaller sample to feature lower MSE scores than the bootstrap sample.

FIG. 2. Relative frequency of occurrence by archetype for JRA55 (circles) and lead day 0 ECWMF-S2S hindcasts (crosses) between 1997 and 2016. Statistics are shown for all cases (blue) and for confident cases (orange). Relative frequencies of occurrence for all other lead times are almost identical to those shown here for lead day 0.

To test this, we apply the one-tailed Mann–Whitney U test (Mann and Whitney 1947) with $\alpha = 0.05$ to test whether the archetype-derived distribution is stochastically smaller than the archetype-independent distribution. A statistically significant result is interpreted as the archetype framework having a positive effect on precipitation hindcast skill.

3. Results

a. Archetype definitions, frequencies of occurrence, and relationship with observed precipitation

The six archetypes (rightmost column of Fig. 1) each exhibit one synoptic wavelength over the domain. Archetype 1 is a zonal pattern, and of the six archetypes it modifies the climatological mean Z_{500} field the least, with a longwave trough situated to the southwest of the landmass and a ridge off the east coast (Pook et al. 2006, 2012, 2014). Correspondingly, this archetype is the most frequently occurring in JRA55 ($\sim 35\%$ of days; Fig. 2) and is associated with precipitation close to the climatological average for all three regions (Figs. 1a–c).

Archetypes 2 and 3 are somewhat related patterns, and occur with the same frequency ($\sim 22\%$ of days; Fig. 2). Both feature strong positive Z_{500} anomalies in the Southern Ocean which, due to their position at relatively high latitudes, can be interpreted as blocks. As shown by the full field contours, these blocks disrupt the climatological westerly flow that is crucial for precipitation in western Tasmania. It is therefore unsurprising that precipitation in this region is lower than average when these archetypes occur (Figs. 1f,i). The block in archetype 3 would interrupt the frontal systems that move over southwest Western Australia and are the main sources of precipitation in this region (Pook et al. 2012, 2014). That the precipitation distribution is virtually unchanged from climatology under this archetype (Fig. 1h) implies that the majority of precipitation on these days occurred as a result of cutoff lows, which contribute $\sim 30\%$ of the region's

precipitation, compared to $\sim 50\%$ from fronts (Pook et al. 2012, 2014).

The negative Z_{500} anomaly over southwest Australia in archetype 4 sits on top of the climatological trough, deepening the low and helping to explain the above average precipitation in southwest Western Australia during occurrences of this archetype (Fig. 1k). Furthermore, the high pressure anomalies to the south of the domain and low pressure anomalies over the midlatitudes shift the westerly storm track further north. This resembles the negative phase of the southern annular mode (SAM; see Thompson and Wallace 2000, and references therein), which is associated with increased austral winter precipitation for parts of Australia, including in the southwest (Hendon et al. 2007). This aligns with the southwest Western Australia precipitation anomalies associated with this archetype (Fig. 1k). However, precipitation for western Tasmania is lower than climatology for this archetype (Fig. 1l), which is not a feature of a negative SAM. The negative precipitation anomalies in this case can be explained by the ridge to the west of Tasmania, diverting flow southward of the state.

Archetype 5 features a ridge over southwest Australia and a trough over the southeast, in what is essentially a reversal of the climatological mean. Unsurprisingly, then, this archetype is the least common ($\sim 5\%$ of days; Fig. 2) and is associated with some of the greatest precipitation anomalies, particularly in western Tasmania due to favorable conditions for enhanced westerly flow over the region (Figs. 1m–o). The northward shift of the storm track over the southeast of mainland Australia helps to explain positive precipitation anomalies for the Murray Basin, while the southwestern ridge prevents precipitation-bearing weather systems reaching southwest Western Australia.

Archetype 6 features a block to the southwest of New Zealand, which results in a pinched trough that extends far north into Australia. This trough resembles a cutoff low (Pook et al. 2006) and forces flow over all three regions, leading to wetter than average conditions, particularly for the Murray Basin and western Tasmania (Figs. 1p–r).

Restricting the analysis to confident periods in JRA55 yields archetype-conditional precipitation distributions that generally differ to the unconditional case. Precipitation anomalies when considering all days of assigned archetypes (i.e., no imposed SSD threshold; blue CDFs in Fig. 1) are of greater magnitude when we subsample confident occurrences (orange CDFs in Fig. 1). This is a useful result as the extent to which CPs can distinguish different precipitation patterns can be crucial—too little distinction and the CPs are not useful for estimating precipitation (Richardson et al. 2018).

As mentioned in the archetype descriptions previously, the archetypes occur with different frequencies. Archetype 1 is the most common archetype in JRA55 between 1997 and 2016 (i.e., the years corresponding to the hindcast data), occurring 7 times more often than the least common archetype (archetype 5; Fig. 2). The differences in frequencies of occurrences look the same when the entire JRA55 period is considered (1958–2016; Fig. 1). There is little seasonal variation in the archetypes' frequencies of occurrence; the largest difference is for archetype 1, which occurs ~ 1.2 times more often during austral

FIG. 3. Proportion of ECMWF-S2S confident cases (i.e., across initialization date and ensemble member) by lead time day and archetype.

summer compared to winter (not shown). Generating the archetypes on individual seasons yields a similar set of archetypes compared to the annual case here, with at least five of the six archetypes in each season closely resembling one of the annual archetypes (see the supplemental material). When considering the subset of high confidence archetypes, how frequently the archetypes occur relative to each other does not change (e.g., archetype 1 is still the most common and archetype 5 is still the least common). This has the consequence of low sample sizes for archetypes 4–6, at 16 (115), 16 (36), and 31 (103) days, respectively, for JRA55 between 1997 and 2016 (1958 and 2016) for the confident subsample (Fig. 2).

Note that the fact that AA generates CPs that represent the data “extremes” does not preclude the CPs from being persistent. We find that archetype 1 persists for more than one day over 70% of the time, compared to $\sim 65\%$ for archetypes 2 and 3, $\sim 50\%$ for archetypes 4, 5, and 6. Furthermore, it is not unusual for archetypes 1, 2, and 3 (4, 5, and 6) to persist for up to 10 (5) days (not shown).

b. Hindcast archetype behavior

The frequencies of hindcast archetypes at a lead time of 0 days are similar to those from JRA55, although the hindcast model overpredicts archetypes 4–6 at the expense of archetypes 1–3 (Fig. 2). These are minor differences, however, implying that the initialization states of ECMWF-S2S look similar to the JRA55 reanalysis, at least for Z_{500} . Although the relative frequencies are largely equivalent, the sample sizes of archetypes 4–6 are not as low in the hindcast as for JRA55 due to multiple ensemble members (66, 55, and 101 days, respectively).

The similarity between JRA55 and ECMWF-S2S Z_{500} fields is evident at all lead times, as shown by the proportion of

FIG. 4. Relative frequency of the number of unique archetypes predicted, as a function of the number of concurrently (i.e., simultaneously) confident ensemble members. Subplots are for different lead times (in days). The number of unique archetypes is bounded by the number of concurrent confident members. For example, when two members are confident (at $x = 2$ for each panel), there is a maximum of two unique archetypes. The maximum number of unique archetypes in the figure is four [in (h), (i), and (j), with very small relative frequencies].

confident days being almost constant with lead time (Fig. 3). The average SSD between the archetype definitions and the underlying hindcast fields is also constant with lead time (not shown). There is a suggestion that archetype 2 makes up a greater proportion of the confident sample at longer leads compared to shorter leads at the expense of archetype 1, but these differences are minor and likely insignificant.

There is, however, a relationship between lead time and the number of ensemble members concurrently (i.e., simultaneously) predicting an archetype with high confidence. At a lead of 0 days the ensemble members differ by initialization state alone; there is no forecast over which the members evolve. Therefore, as expected, on the vast majority of occasions (over 80%) that any member is confident at lead day 0, all members are confident. Furthermore, these members always predict the same archetype (Fig. 4a). As the lead time increases, it becomes more common for smaller numbers of members to be confident at the same time. By a lead time of 4 days (Fig. 4e), the distribution flips and the modal case is for one confident member rather than 11 members. The ensemble members tend to predict the same archetype for all lead times up to 1 week (Figs. 4a–f). Beyond this, members diverge sufficiently to allow for multiple archetypes predicted with high confidence on the same day (Figs. 4g–j). As the archetypes are distinct from each other in terms of their Z_{500} anomaly patterns (Fig. 1), such cases are unlikely to be helpful in identifying

periods of high forecast confidence as they imply significant divergence in the ensemble members. The most promising cases, then, are those occasions for which multiple members are confidently predicting the same archetype. Beyond lead times of several days, these situations are relatively rare, particularly for when more than three members are simultaneously confident.

c. Precipitation hindcast skill during periods of high confidence

For brevity, we shall henceforth use the phrase “greater skill” (“lower skill”) to explicitly mean a more (less) accurate precipitation hindcast, as measured by the MSE, for those days corresponding to an archetype predicted with high confidence, compared to other days, i.e., Eqs. (3) and (10). We first discuss results comparing high confidence to normal periods (the blue lines and shading in Fig. 5), and will follow this with the differences between these results and those comparing high and low confidence.

Overall, there is greater skill for short leads in Murray Basin and western Tasmania (up to 6 and 11 days, respectively) and for longer leads in southwest Western Australia (over 6 days) (Figs. 5s–u). This statement is based on the median (of the ensemble members’ independently calculated MSE scores), which means that the model has greater skill for the stated regions and leads on over 50% of confident forecast occasions.

FIG. 5. (a)–(u) Precipitation hindcast skill for periods of high model confidence in the predicted archetype relative to periods of normal (blue) and low (orange) confidence. The data are the ratios of the mean squared error for confident periods to normal/low confidence periods; see the text for details. The solid blue and orange lines are the ensemble medians and the shading represents the 90% confidence intervals. Data below the black line correspond to greater skill during confident periods, while data above the black line correspond to lower skill. The black line in (a)–(r) also represents the ensemble median sample size of each archetype as a fraction of the total sample size [i.e., Eq. (10)]. The first six rows show results for each archetype [the r_k in Eq. (8)], and the results for all archetypes are shown in the bottom row [Eq. (3)]. The insets zoom in on their parent panels.

Indeed, the median lies within the area of greater skill (i.e., below the black line) for almost every lead time for both the Murray Basin and southwest Western Australia. However, in all cases the confidence intervals include both greater and lower skill except for leads of 1, 2, 4, and 5 days in western Tasmania, for which they lie within the area indicating greater skill. These results suggest that the majority of the time we can expect greater precipitation forecast accuracy when the archetypes are predicted with high confidence for ≥ 1 week leads in southwest Western Australia and for ≤ 1 week leads in the Murray Basin and western Tasmania. Otherwise, high confidence in a predicted archetype yields no additional value in assessing whether we can expect greater skill.

The potential utility of using archetypes to identify situations of greater precipitation skill is further analyzed by comparing skill during high confidence in the predicted archetype (the lower 10% of the Z_{500} anomaly field-to-archetype SSD distribution) with skill during low confidence (the upper 10% of the SSD distribution). If the archetypes have any role in distinguishing periods of improved skill, then this comparison (orange lines and shading in Fig. 5) is expected to emphasize the potentially useful forecast situations identified by previously described results (i.e., comparisons between high and normal confidence; blue lines and shading in Fig. 5).

We find that, in general, this hypothesis is borne out. Skill comparisons between high and low confidence is qualitatively similar to that comparing high and normal confidence, but with some important differences. Median scores for the former are sometimes distinctly more skillful than for the latter. Such behavior occurs for the cases in which we observe greater skill, namely, shorter lead times in western Tasmania and the Murray Basin, and longer lead times in southwest Western Australia (cf. the blue and orange solid lines in Figs. 5s–u). However, as the medians of both comparisons tend to lie within the confidence bounds of the other method, these differences are likely insignificant.

The major difference between the two sets of results is the distinct narrowing of the confidence intervals, a feature that is evident for almost all archetypes, regions and lead times (Fig. 5). In some cases, this has a large impact on interpreting the significance of the results. Consider the overall results for western Tasmania (Fig. 5u). The narrower confidence intervals (orange shading) suggest that the method is able to identify periods of greater skill for the first 10 days of the hindcast, and that these results are statistically significant. When comparing high confidence with normal confidence, however, this statement only applies to lead times of 1, 2, 4, and 5 days (blue shading).

Furthermore, statistically significant results of greater skill in western Tasmania for archetypes 1 and 3 are apparent for the first 7 days of the hindcast (orange shading; Figs. 5c,i), compared to the first 4 days when comparing high confidence with normal confidence (blue shading). For archetype 2 the increase is from 7 to 14 days (Fig. 5f). For southwest Western Australia, the biggest change in the results is for archetype 1 (Fig. 5b), which displays statistically significant greater skill consistently

for lead times of more than 30 days (orange shading). This affects the corresponding overall results (Fig. 5t), bringing the upper confidence bound much closer to the area of greater skill.

The fact that the skill tends to increase for comparisons of high and low confidence compared to those for high and normal confidence (at least in situations where high versus normal was somewhat skillful anyway) suggests that there is some relationship between a high confidence archetype prediction and the likely precipitation hindcast skill associated with it.

d. Relationship between skill and “dry” archetypes

Figure 5 implies a relationship between skill and the precipitation associated with each archetype, with “dry” archetypes typically yielding greater skill than “wet” archetypes. For western Tasmania, archetypes 2 and 3 feature greater skill (i.e., low r_k) for lead times of up to 2 weeks (Figs. 5f,i), and these are the driest archetypes in this region (Figs. 1f,i). Also note how the confidence intervals for these cases lie within the area of greater skill for up to 7 days lead (i.e., are statistically significant), a feature that is rare in these results. It is encouraging that the gain in skill during confident periods for these two archetypes, plus for archetype 1 (Fig. 5c), is for the three most common archetypes, which together account for the vast majority of all archetype occurrences (Fig. 2). This implies that we would be able to exploit a forecast of greater skill relatively often.

Furthermore, there is some evidence that for more confident hindcasts (i.e., for greater numbers of ensemble members that simultaneously predict an archetype with confidence), the corresponding precipitation hindcast mean and standard deviation decreases (Fig. 6). For a 1-day lead, the vast majority of days feature either 0 or all 11 members being confident, with a clear reduction in the upper tail of the ensemble mean precipitation distribution for the latter case (Fig. 6a). Similarly, for lead times of 10 and 30 days, there is a steady decrease in mean precipitation for higher numbers of confident members (Figs. 6c,d). In all cases, higher percentiles of the precipitation ensemble mean distribution decrease faster than lower percentiles, most clearly shown by the values over the 95th percentile (i.e., the orange markers that lie above the boxplots’ uppermost whiskers).

From Figs. 5 and 6 we are not able to conclude whether the greater skill is a result of the dry archetype or due to the fact that predictions of dry days tend to be more accurate regardless (i.e., independent of the archetype). Figure 7 provides results to test this, as described in section 2f. For up to nearly 2-weeks lead time, confident predictions of archetype 2 yield more accurate western Tasmania precipitation hindcasts compared to those on dry days alone. This is evidenced by generally lower MSE results for each ensemble member compared to random sample MSE results for dry days, and corresponding statistically significant Mann–Whitney U test results. This shows that confidence in an upcoming occurrence archetype 2 is of benefit to predicting precipitation in western Tasmania. The hindcast error appears to saturate after a 2-week lead time, mirroring the results comparing skill during confident periods with skill

FIG. 6. Western Tasmania precipitation hindcast (top) ensemble mean and (bottom) standard deviation as a function of the number of ensemble members that concurrently predict an archetype with high confidence. Orange markers represent all the data. Boxplots are only drawn when the sample size is at least 20. The boxes show the 25th, 50th, and 75th percentiles, and the whiskers are at the 5th and 95th percentiles.

during normal/low confidence periods (Fig. 5f). Similar results to those presented in Fig. 7 are evident for archetypes 1 and 3, with statistical significance up to lead times of 7 and 4 days, respectively (not shown).

While there is a link between greater skill and dry archetypes in southwest Western Australia and the Murray Basin, the greater skill is more likely due to the overall higher accuracy of dry day hindcasts. Archetypes 1 and (especially) 5 are typically dry for southwest Western Australia (Figs. 1b,n), and are associated with greater skill (Figs. 5b,n). The skill for this region during confident predictions of archetype 5 is extremely good, with almost 0 error at all lead times, as shown by the solid lines in Fig. 1n being very close to 0. This is almost certainly due to the model predicting dry days accurately: when the model is confident about archetype 5 occurring in the forecast, the precipitation predicted and the resultant observation will be close to 0. The two driest archetypes for the Murray Basin (1 and 3; Figs. 1a,g) exhibit greater median skill (solid lines in Figs. 5a,g). However, so does archetype 5 (Fig. 5m), which is associated with wetter than climatological conditions (Fig. 1m). For both southwest Western Australia and the Murray Basin, there is little difference between the skill on confident dry archetype days and on a random sample of dry

hindcast days (not shown), implying that precipitation hindcast skill for these regions is not related to the archetypes.

4. Discussion and conclusions

We have introduced a new methodology for identifying periods of high confidence in subseasonal forecasts using a CP-based approach, and showed that this method can identify days of improved precipitation forecast skill for southern Australia. CPs were generated with Z_{500} anomaly fields over Australia using archetypal analysis, which attempts to capture the extremes of the data. These archetypes are varying combinations of ridges and troughs, and have different relationships with precipitation depending on the locations of these features. By subsampling archetype occurrences according to how well matched they are to the underlying reanalysis fields, we showed that their precipitation distributions could be modified. This achieved greater distinction between archetypes, as precipitation anomalies associated with particular archetypes were magnified after subsampling i.e., wet archetypes became wetter and dry archetypes became drier.

We used a 20-yr ECMWF-S2S hindcast dataset to derive hindcasts of the archetypes, which we subsampled to identify occasions for which the model was particularly confident about

FIG. 7. Western Tasmania precipitation hindcast mean squared error (MSE) for confident hindcasts of archetype 2 for each ensemble member (orange circles). Boxplots show 11 000 bootstrapped MSE results (1000 samples per ensemble member) for dry day hindcasts, where a dry day is any hindcast precipitation amount equal to or less than the maximum precipitation hindcast value that coincides with a confident prediction of archetype 2 at that lead time. The boxes show the 25th, 50th, and 75th percentiles, and the whiskers are at the 5th and 95th percentiles. Black squares indicate statistical significance of a one-tailed Mann-Whitney U test. Only the first 14 lead time days are shown; no other lead times yield significant results.

the likelihood of predicted archetypes. We then assessed the skill in hindcasting precipitation for three regions in southern Australia on these days of high confidence compared to normal and compared to days of low confidence. Results for all archetypes showed that skill during these confident periods was greater than normal for lead times of less than 6 days in Murray Basin, less than 11 days in western Tasmania, and greater than 6 days in southwest Western Australia. However, confidence intervals suggested this skill improvement across all archetypes was not significant in almost every case.

By breaking down the skill for each archetype individually, we showed that those archetypes typically associated with drier-than-average conditions had greater skill than wetter-than-average archetypes. However, this greater skill only appears to be related to the archetype framework for western Tasmania. For the other two regions, similarly accurate hindcasts can be obtained by subsampling dry days from the hindcast, independent of the archetype.

When comparing skill during periods of high and low confidence, there was general improvement in median skill over the results comparing high and normal confidence periods. This is particularly for those cases that already exhibited greater skill compared to normal. Moreover, these results were far more likely to be statistically significant due to much narrower confidence bounds. This comparison of high versus low confidence is effectively a test of how well the method distinguishes precipitation hindcast skill. When the model hindcast fields are very similar to the archetypes, the precipitation hindcast accuracy is much greater than when the model hindcast fields are dissimilar to any archetype. This suggests the method is highlighting a relationship between Z_{500} and precipitation that is predictable, even for lead times of over 1 month in southwest Western Australia (Fig. 5b).

The stricter test of the method is when comparing the high with normal confidence case, which is essentially testing how

useful the method is compared to the usual forecast framework in which we do not consider archetypes. Here, due to the uncertainty bounds on the skill results, the potential usefulness of the method is unclear, as there only a few situations in which we could be sure that the method is identifying periods of greater skill.

A crucial point is that the skill results presented are particular to the set of archetypes generated. As discussed in section 2a, there were several decisions made such as the domain used, spatial resolution, and the number of archetypes to be generated. We tested classifications of 4, 6, and 12 archetypes. While the different combinations yielded qualitatively the same skill, the set of 6 resulted in the greatest skill gain during confident periods, in particular for short lead times in Murray Basin and western Tasmania, which were not as clearly skillful as with 6 archetypes. Given the potential for different results, and in line with other studies (e.g., Neal et al. 2016; Kučerová et al. 2017), we urge users of CP classifications to test several variants to ascertain the most promising classification for their application.

An interesting piece of future analysis would be to extend this methodology from a deterministic to a probabilistic framework. In the work presented here, for the shortest leads (no more than a few days), a confident forecast is normally characterized by all or most ensemble members being confident about a particular archetype occurring (e.g., Fig. 4a). In these cases, it is possible to shift to probabilistic verification measures as there are typically enough data points available for their calculation. For longer lead times, there are often only one or two ensemble members that are confident about the predicted archetype for any given forecast, leaving too few data points from which to calculate probabilistic verification scores. One solution could be to retain all ensemble members, but weight them according to how confident they are, as opposed to a binary keep/reject based on a confidence threshold. Another solution might be to make use of the real-time ECMWF-S2S dataset, which has 51 ensemble members. This larger ensemble size would hopefully yield enough confident predictions of archetypes at longer lead times on which to calculate probabilistic verification measures.

Acknowledgments. The authors thank three reviewers for their comments on this article. The authors received funding for this work from the CSIRO Decadal Climate Forecasting Project and the Digiscape Future Science Platform. The authors declare no conflicts of interest.

Data availability statement. The JRA55 data were provided by the Japan Meteorological Agency (2013), which were accessed from <https://rda.ucar.edu/datasets/ds628.0/>. ECMWF S2S hindcast data were downloaded from <https://apps.ecmwf.int/datasets/data/s2s-reforecasts-instantaneous-accum-ecmf/>. The AWAP data were provided by Peter Briggs. The shape files for the natural resource management regions were obtained from <https://www.climatechangeinaustralia.gov.au/en/climate-projections/about/modelling-choices-and-methodology/regionalisation-schemes/>.

The archetypes were generated using the Matlab package Principal Convex Hull/Archetypal Analysis available at

http://www.mortenmorup.dk/MMhomepageUpdated_files/Page327.htm. All other analyses were conducted using Python 3 with the following packages: Cartopy (Met Office 2015), Dask (Dask Development Team 2016), Matplotlib (Hunter 2007), pandas (McKinney 2010), NumPy (van der Walt et al. 2011), and xarray (Hoyer and Hamman 2017).

REFERENCES

- Bennett, J. C., M. R. Grose, S. P. Corney, C. J. White, G. K. Holz, J. J. Katzfey, D. A. Post, and N. L. Bindoff, 2014: Performance of an empirical bias-correction of a high-resolution climate dataset. *Int. J. Climatol.*, **34**, 2189–2204, <https://doi.org/10.1002/joc.3830>.
- Bogardi, I., I. Matyasovszky, A. Bardossy, and L. Duckstein, 1993: Application of a space-time stochastic model for daily precipitation using atmospheric circulation patterns. *J. Geophys. Res.*, **98**, 16 653–16 667, <https://doi.org/10.1029/JD00919>.
- Boumal, N., B. Mishra, P.-A. Absil, and R. Sepulchre, 2014: Manopt, a Matlab toolbox for optimization on manifolds. *J. Mach. Learn. Res.*, **15**, 1455–1459, <http://jmlr.org/papers/v15/boumal14a.html>.
- Casado, M., M. Pastor, and F. Doblas-Reyes, 2010: Links between circulation types and precipitation over Spain. *Phys. Chem. Earth*, **35**, 437–447, <https://doi.org/10.1016/j.pce.2009.12.007>.
- Christiansen, B., 2007: Atmospheric circulation regimes: Can cluster analysis provide the number? *J. Climate*, **20**, 2229–2250, <https://doi.org/10.1175/JCLI4107.1>.
- CSIRO and Bureau of Meteorology, 2015: Climate Change in Australia website. <https://www.climatechangeinaustralia.gov.au/en/climate-projections/about/modelling-choices-and-methodology/regionalisation-schemes/>.
- Cutler, A., and L. Breiman, 1994: Archetypal analysis. *Technometrics*, **36**, 338–347, <https://doi.org/10.1080/00401706.1994.10485840>.
- Dalcher, A., E. Kalnay, and R. N. Hoffman, 1988: Medium range lagged average forecasts. *Mon. Wea. Rev.*, **116**, 402–416, [https://doi.org/10.1175/1520-0493\(1988\)116<0402:MRLAF>2.0.CO;2](https://doi.org/10.1175/1520-0493(1988)116<0402:MRLAF>2.0.CO;2).
- Dask Development Team, 2016: Dask: Library for dynamic task scheduling. <https://dask.org>.
- de Andrade, F. M., C. A. S. Coelho, and I. F. A. Cavalcanti, 2019: Global precipitation hindcast quality assessment of the Subseasonal to Seasonal (S2S) prediction project models. *Climate Dyn.*, **52**, 5451–5475, <https://doi.org/10.1007/s00382-018-4457-z>.
- Fereday, D. R., J. R. Knight, A. A. Scaife, C. K. Folland, and A. Philipp, 2008: Cluster analysis of North Atlantic/European circulation types and links with tropical Pacific sea surface temperatures. *J. Climate*, **21**, 3687–3703, <https://doi.org/10.1175/2007JCLI1875.1>.
- Ferranti, L., S. Corti, and M. Janousek, 2015: Flow-dependent verification of the ECMWF ensemble over the Euro-Atlantic sector. *Quart. J. Roy. Meteor. Soc.*, **141**, 916–924, <https://doi.org/10.1002/qj.2411>.
- , L. Magnusson, F. Vitart, and D. S. Richardson, 2018: How far in advance can we predict changes in large-scale flow leading to severe cold conditions over Europe? *Quart. J. Roy. Meteor. Soc.*, **144**, 1788–1802, <https://doi.org/10.1002/qj.3341>.
- Frame, T. H. A., J. Methven, S. L. Gray, and M. H. P. Ambaum, 2013: Flow-dependent predictability of the North Atlantic jet. *Geophys. Res. Lett.*, **40**, 2411–2416, <https://doi.org/10.1002/grl.50454>.
- Frías, M. D., S. Herrera, A. S. Cofiño, and J. M. Gutiérrez, 2010: Assessing the skill of precipitation and temperature seasonal forecasts in Spain: Windows of opportunity related to ENSO events. *J. Climate*, **23**, 209–220, <https://doi.org/10.1175/2009JCLI2824.1>.
- Gerstengarbe, F. W., and P. C. Werner, 1997: A method to estimate the statistical confidence of cluster separation. *Theor. Appl. Climatol.*, **57**, 103–110, <https://doi.org/10.1007/BF00867981>.
- Goddard, L., and M. Dille, 2005: El Niño: Catastrophe or opportunity. *J. Climate*, **18**, 651–665, <https://doi.org/10.1175/JCLI-3277.1>.
- Gudmundsson, L., J. B. Bremnes, J. E. Haugen, and T. Engen-Skaugen, 2012: Technical Note: Downscaling RCM precipitation to the station scale using statistical transformations: a comparison of methods. *Hydrol. Earth Syst. Sci.*, **16**, 3383–3390, <https://doi.org/10.5194/hess-16-3383-2012>.
- Hannachi, A., and N. Trendafilov, 2017: Archetypal analysis: Mining weather and climate extremes. *J. Climate*, **30**, 6927–6944, <https://doi.org/10.1175/JCLI-D-16-0798.1>.
- , D. M. Straus, C. L. E. Franzke, S. Corti, and T. Woollings, 2017: Low-frequency nonlinearity and regime behavior in the Northern Hemisphere extratropical atmosphere. *Rev. Geophys.*, **55**, 199–234, <https://doi.org/10.1002/2015RG000509>.
- Hendon, H. H., D. W. J. Thompson, and M. C. Wheeler, 2007: Australian rainfall and surface temperature variations associated with the Southern Hemisphere annular mode. *J. Climate*, **20**, 2452–2467, <https://doi.org/10.1175/JCLI4134.1>.
- Hoskins, B., 2013: The potential for skill across the range of the seamless weather-climate prediction problem: A stimulus for our science. *Quart. J. Roy. Meteor. Soc.*, **139**, 573–584, <https://doi.org/10.1002/qj.1991>.
- Hoyer, S., and J. Hamman, 2017: xarray: N-D labeled Arrays and Datasets in Python. *J. Open Res. Softw.*, **5**, 10, <https://doi.org/10.5334/jors.148>.
- Hunter, J. D., 2007: Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.*, **9**, 90–95, <https://doi.org/10.1109/MCSE.2007.55>.
- Huth, R., C. Beck, A. Philipp, M. Demuzere, Z. Ustrnul, M. Cahynová, J. Kyseľ, and O. E. Tveito, 2008: Classifications of atmospheric circulation patterns. *Ann. N. Y. Acad. Sci.*, **1146**, 105–152, <https://doi.org/10.1196/annals.1446.019>.
- Japan Meteorological Agency, 2013: JRA-55: Japanese 55-year Reanalysis, Daily 3-Hourly and 6-Hourly Data. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory, accessed 10 November 2020, <https://doi.org/10.5065/D6HH6H41>.
- Jenkinson, A. F., and F. P. Collinson, 1977: An initial climatology of gales over the North Sea. *Synoptic Climatology Branch Memo*, 62, 18 pp.
- Jones, C., D. E. Waliser, K. M. Lau, and W. Stern, 2004: The Madden–Julian oscillation and its impact on northern hemisphere weather predictability. *Mon. Wea. Rev.*, **132**, 1462–1471, [https://doi.org/10.1175/1520-0493\(2004\)132<1462:TMOAI>2.0.CO;2](https://doi.org/10.1175/1520-0493(2004)132<1462:TMOAI>2.0.CO;2).
- , A. Hazra, and L. M. V. Carvalho, 2015: The Madden-Julian Oscillation and boreal winter forecast skill: An analysis of NCEP CFSv2 reforecasts. *J. Climate*, **28**, 6297–6307, <https://doi.org/10.1175/JCLI-D-15-0149.1>.
- Jones, D., 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>.
- Kalnay, E., 2019: Historical perspective: Earlier ensembles and forecasting forecast skill. *Quart. J. Roy. Meteor. Soc.*, **145**, 25–34, <https://doi.org/10.1002/qj.3595>.
- , and A. Dalcher, 1987: Forecasting forecast skill. *Mon. Wea. Rev.*, **115**, 349–356, [https://doi.org/10.1175/1520-0493\(1987\)115<0349:FFS>2.0.CO;2](https://doi.org/10.1175/1520-0493(1987)115<0349:FFS>2.0.CO;2).

- Kim, H.-M., P. J. Webster, and J. A. Curry, 2012: Seasonal prediction skill of ECMWF System 4 and NCEP CFSv2 retrospective forecast for the Northern Hemisphere Winter. *Climate Dyn.*, **39**, 2957–2973, <https://doi.org/10.1007/s00382-012-1364-6>.
- Kohonen, T., 2001: *The Basic SOM. Self-Organizing Maps*, T. Kohonen, Ed., Springer, 105–176, https://doi.org/10.1007/978-3-642-56927-2_3.
- Kučerová, M., C. Beck, A. Philipp, and R. Huth, 2017: Trends in frequency and persistence of atmospheric circulation types over Europe derived from a multitude of classifications. *Int. J. Climatol.*, **37**, 2502–2521, <https://doi.org/10.1002/joc.4861>.
- Lamb, H. H., 1972: *British Isles Weather Types and a Register of Daily Sequence of Circulation Patterns 1861–1971. Geophysical Memoirs*, Vol. 116, H.M. Stationery Office, 85 pp.
- Lavaysse, C., J. Vogt, A. Toreti, M. L. Carrera, and F. Pappenberger, 2018: On the use of weather regimes to forecast meteorological drought over Europe. *Nat. Hazards Earth Syst. Sci.*, **18**, 3297–3309, <https://doi.org/10.5194/nhess-18-3297-2018>.
- Li, S., and A. W. Robertson, 2015: Evaluation of submonthly precipitation forecast skill from global ensemble prediction systems. *Mon. Wea. Rev.*, **143**, 2871–2889, <https://doi.org/10.1175/MWR-D-14-00277.1>.
- Mann, H. B., and D. R. Whitney, 1947: On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.*, **18**, 50–60.
- Manzanas, R., M. D. Frías, A. S. Cofiño, and J. M. Gutiérrez, 2014: Validation of 40 year multimodel seasonal precipitation forecasts: The role of ENSO on the global skill. *J. Geophys. Res. Atmos.*, **119**, 1708–1719, <https://doi.org/10.1002/2013JD020680>.
- Mariotti, A., and Coauthors, 2020: Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond. *Bull. Amer. Meteor. Soc.*, **101**, E608–E625, <https://doi.org/10.1175/BAMS-D-18-0326.1>.
- McKinney, W., 2010: Data structures for statistical computing in Python. *Proceedings 9th Python in Science Conference*, S. van der Walt and J. Millman, Eds., 51–56, SciPy, <https://doi.org/10.25080/Majora-92bf1922-012>.
- Met Office, 2015: Cartopy: A cartographic Python library with a matplotlib interface. <http://scitools.org.uk/cartopy>.
- Michelangeli, P.-A., R. Vautard, and B. Legras, 1995: Weather regimes: Recurrence and quasi stationarity. *J. Atmos. Sci.*, **52**, 1237–1256, [https://doi.org/10.1175/1520-0469\(1995\)052<1237:WRRAQ5>2.0.CO;2](https://doi.org/10.1175/1520-0469(1995)052<1237:WRRAQ5>2.0.CO;2).
- Miller, D. E., and Z. Wang, 2019: Assessing seasonal predictability sources and windows of high predictability in the climate forecast system, version 2. *J. Climate*, **32**, 1307–1326, <https://doi.org/10.1175/JCLI-D-18-0389.1>.
- Milligan, G. W., and M. C. Cooper, 1985: An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, **50**, 159–179, <https://doi.org/10.1007/BF02294245>.
- Molteni, F., and T. N. Palmer, 1991: A real-time scheme for the prediction of forecast skill. *Mon. Wea. Rev.*, **119**, 1088–1097, [https://doi.org/10.1175/1520-0493\(1991\)119<1088:ARTSFT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1991)119<1088:ARTSFT>2.0.CO;2).
- Mørup, M., and L. K. Hansen, 2012: Archetypal analysis for machine learning and data mining. *Neurocomputing*, **80**, 54–63, <https://doi.org/10.1016/j.neucom.2011.06.033>.
- Muñoz, G. L., S. J. Goddard, Mason, and A. W. Robertson, 2016: Cross-time scale interactions and rainfall extreme events in southeastern South America for the Austral summer. Part II: Predictive skill. *J. Climate*, **29**, 5915–5934, <https://doi.org/10.1175/JCLI-D-15-0699.1>.
- Neal, R., D. Fereday, R. Crocker, and R. E. Comer, 2016: A flexible approach to defining weather patterns and their application in weather forecasting over Europe. *Meteor. Appl.*, **23**, 389–400, <https://doi.org/10.1002/met.1563>.
- , R. Dankers, A. Saulter, A. Lane, J. Millard, G. Robbins, and D. Price, 2018: Use of probabilistic medium- to long-range weather-pattern forecasts for identifying periods with an increased likelihood of coastal flooding around the UK. *Meteor. Appl.*, **25**, 534–547, <https://doi.org/10.1002/met.1719>.
- Neena, J. M., J. Y. Lee, D. Waliser, B. Wang, and X. Jiang, 2014: Predictability of the Madden–Julian oscillation in the Intraseasonal Variability Hindcast Experiment (ISVHE). *J. Climate*, **27**, 4531–4543, <https://doi.org/10.1175/JCLI-D-13-00624.1>.
- Nigro, M. A., J. J. Cassano, and M. W. Seefeldt, 2011: A weather-pattern-based approach to evaluate the Antarctic Mesoscale Prediction System (AMPS) forecasts: Comparison to automatic weather station observations. *Wea. Forecasting*, **26**, 184–198, <https://doi.org/10.1175/2010WAF2222444.1>.
- Palmer, T. N., and S. Tibaldi, 1988: On the prediction of forecast skill. *Mon. Wea. Rev.*, **116**, 2453–2480, [https://doi.org/10.1175/1520-0493\(1988\)116<2453:OTPOFS>2.0.CO;2](https://doi.org/10.1175/1520-0493(1988)116<2453:OTPOFS>2.0.CO;2).
- Philipp, A., P. M. Della-Marta, J. Jacobeit, D. R. Fereday, P. D. Jones, A. Moberg, and H. Wanner, 2007: Long-term variability of daily North Atlantic European pressure patterns since 1850 classified by simulated annealing clustering. *J. Climate*, **20**, 4065–4095, <https://doi.org/10.1175/JCLI4175.1>.
- Piani, C., J. O. Haerter, and E. Coppola, 2010: Statistical bias correction for daily precipitation in regional climate models over Europe. *Theor. Appl. Climatol.*, **99**, 187–192, <https://doi.org/10.1007/s00704-009-0134-9>.
- Pook, M. J., P. C. McIntosh, and G. A. Meyers, 2006: The synoptic decomposition of cool-season rainfall in the southeastern Australian cropping region. *J. Appl. Meteor. Climatol.*, **45**, 1156–1170, <https://doi.org/10.1175/JAM2394.1>.
- , J. S. Risbey, and P. C. McIntosh, 2012: The synoptic climatology of cool-season rainfall in the central wheatbelt of Western Australia. *Mon. Wea. Rev.*, **140**, 28–43, <https://doi.org/10.1175/MWR-D-11-00048.1>.
- , —, and —, 2014: A comparative synoptic climatology of cool-season rainfall in major grain-growing regions of southern Australia. *Theor. Appl. Climatol.*, **117**, 521–533, <https://doi.org/10.1007/s00704-013-1021-y>.
- Qin, J., and W. A. Robinson, 1995: The impact of tropical forcing on extratropical predictability in a simple global model. *J. Atmos. Sci.*, **52**, 3895–3910, [https://doi.org/10.1175/1520-0469\(1995\)052<3895:TIOFTFO>2.0.CO;2](https://doi.org/10.1175/1520-0469(1995)052<3895:TIOFTFO>2.0.CO;2).
- Ratri, D. N., K. Whan, and M. Schmeits, 2019: A comparative verification of raw and bias-corrected ECMWF seasonal ensemble precipitation reforecasts in Java (Indonesia). *J. Appl. Meteor. Climatol.*, **58**, 1709–1723, <https://doi.org/10.1175/JAMC-D-18-0210.1>.
- Raupach, M. R., P. R. Briggs, V. Haverd, E. A. King, M. Paget, and C. M. Trudinger, 2009: Australian water availability project (AWAP): CSIRO Marine and Atmospheric Research component: Final report for phase 3. CAWCR Tech. Rep. 013, 67 pp., https://www.cawcr.gov.au/technical-reports/CTR_013.pdf.
- Richardson, D., H. J. Fowler, C. G. Kilsby, and R. Neal, 2018: A new precipitation and drought climatology based on weather patterns. *Int. J. Climatol.*, **38**, 630–648, <https://doi.org/10.1002/joc.5199>.
- , C. G. Kilsby, H. J. Fowler, and A. Bárdossy, 2019: Weekly to multi-month persistence in sets of daily weather patterns over Europe and the North Atlantic Ocean. *Int. J. Climatol.*, **39**, 2041–2056, <https://doi.org/10.1002/joc.5932>.

- , H. J. Fowler, C. G. Kilsby, R. Neal, and R. Dankers, 2020a: Improving sub-seasonal forecast skill of meteorological drought: A weather pattern approach. *Nat. Hazards Earth Syst. Sci.*, **20**, 107–124, <https://doi.org/10.5194/nhess-20-107-2020>.
- , R. Neal, R. Dankers, K. Mylne, R. Cowling, H. Clements, and J. Millard, 2020b: Linking weather patterns to regional extreme precipitation for highlighting potential flood events in medium- to long-range forecasts. *Meteor. Appl.*, **27**, e1931, <https://doi.org/10.1002/met.1931>.
- Rodwell, M. J., and F. J. Doblas-Reyes, 2006: Medium-range, monthly, and seasonal prediction for Europe and the use of forecast information. *J. Climate*, **19**, 6025–6046, <https://doi.org/10.1175/JCLI3944.1>.
- Branković, Č., and T. N. Palmer, 2000: Seasonal skill and predictability of ECMWF PROVOST ensembles. *Quart. J. Roy. Meteor. Soc.*, **126**, 2035–2067, <https://doi.org/10.1256/smsqj.56703>.
- Shukla, J., and Coauthors, 2000: Dynamical seasonal prediction. *Bull. Amer. Meteor. Soc.*, **81**, 2593–2606, [https://doi.org/10.1175/1520-0477\(2000\)081<2593:DSP>2.3.CO;2](https://doi.org/10.1175/1520-0477(2000)081<2593:DSP>2.3.CO;2).
- Steinschneider, S., and U. Lall, 2015: Daily precipitation and tropical moisture exports across the Eastern United States: An application of archetypal analysis to identify spatiotemporal structure. *J. Climate*, **28**, 8585–8602, <https://doi.org/10.1175/JCLI-D-15-0340.1>.
- Stephenson, D. B., A. Hannachi, and A. O'Neill, 2004: On the existence of multiple climate regimes. *Quart. J. Roy. Meteor. Soc.*, **130**, 583–605, <https://doi.org/10.1256/qj.02.146>.
- Thompson, D. W. J., and J. M. Wallace, 2000: Annular modes in the extratropical circulation. Part I: Month-to-Month variability. *J. Climate*, **13**, 1000–1016, [https://doi.org/10.1175/1520-0442\(2000\)013<1000:AMITEC>2.0.CO;2](https://doi.org/10.1175/1520-0442(2000)013<1000:AMITEC>2.0.CO;2).
- van der Walt, S., S. C. Colbert, and G. Varoquaux, 2011: The NumPy array: A structure for efficient numerical computation. *Comput. Sci. Eng.*, **13**, 22–30, <https://doi.org/10.1109/MCSE.2011.37>.
- Vigaud, N., A. Robertson, and M. K. Tippett, 2018: Predictability of recurrent weather regimes over North America during winter from submonthly reforecasts. *Mon. Wea. Rev.*, **146**, 2559–2577, <https://doi.org/10.1175/MWR-D-18-0058.1>.
- 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>.
- Vuillaume, J.-F., and S. Herath, 2017: Improving global rainfall forecasting with a weather type approach in Japan. *Hydrol. Sci. J.*, **62**, 167–181, <https://doi.org/10.1080/02626667.2016.1183165>.
- Wobus, R. L., and E. Kalnay, 1995: Three years of operational prediction of forecast skill at NMC. *Mon. Wea. Rev.*, **123**, 2132–2148, [https://doi.org/10.1175/1520-0493\(1995\)123<2132:TYOOPO>2.0.CO;2](https://doi.org/10.1175/1520-0493(1995)123<2132:TYOOPO>2.0.CO;2).
- Zhao, T., J. C. Bennett, Q. J. Wang, A. Schepen, A. W. Wood, D. E. Robertson, and M.-H. Ramos, 2017: How suitable is quantile mapping for postprocessing GCM precipitation forecasts? *J. Climate*, **30**, 3185–3196, <https://doi.org/10.1175/JCLI-D-16-0652.1>.