Book title: SUB- SEASONAL TO SEASONAL PREDICTION (1st Edition): THE GAP BETWEEN WEATHER AND CLIMATE FORECASTING.
Editors: Andrew W. Robertson and Frederic Vitart.
Paperback ISBN: 9780128117149. **eBook ISBN:** 9780128117156. **Imprint:** Elsevier.
Published Date: 24th October 2018.
Page Count: 585

Chapter 17: Forecast verification for S2S time scales. Pages 337 to 361.

Authors: Caio A. S. Coelho, Barbara Brown, Laurie Wilson, Marion Mittermaier, Barbara Casati.

NON-PRINT ITEMS

Abstract

Forecast verification is a key component of a forecasting system and provides information about forecast quality to model and forecast developers and various users. This chapter provides an overview of methods relevant to sub-seasonal to seasonal (S2S) forecast verification, starting with the definition of forecast goodness and some fundamental forecast quality attributes. Next the factors affecting the design of verification studies are presented. The recognition of uncertainties in observational datasets and the need for care in matching forecasts and observations is also discussed. A large part of the chapter is dedicated to a review of the most common deterministic and probabilistic forecast verification measures and a summary of novel spatial verification methods developed during the last two decades. Types of S2S forecasts and current verification practices are presented. The chapter is concluded with a summary, challenges and recommendations for advancing S2S verification research and practice.

Key Words
Forecast quality assessment; forecast performance attributes; forecast verification methods across time and space scales; seamless verification; user-oriented verification

Chapter starts here

1. Introduction

Forecast verification (or evaluation) is a critical aspect in the forecast improvement process, and is also fundamental to inform forecast users regarding their reliability, skill, accuracy, and other features, to aid optimal use. The idea of evaluating forecasts and projections using quantitative methods dates back more than a century and many measures commonly used today for assessing sub-seasonal and seasonal forecasts were developed early in the 20th century for weather forecasts (Murphy 1996). However, several new measures and approaches have been developed in the last couple of decades in response to newly identified needs for different kinds of information, changes in forecast types, and the need to adequately address certain forecast performance questions. For example, spatial methods have become a part of the verification toolbox only in the past 20 years. Verification science continues as an active research area as new forecasts, such as sub-seasonal, are developed and new challenges are discovered. Sub-seasonal forecast verification therefore capitalizes on methodological developments on other time scales (e.g weather and seasonal).

As defined by Murphy (1993), forecast "goodness" combines forecast quality, consistency, and value. Forecast verification, by definition, measures forecast *quality* through comparisons of forecasts to observations. Although forecast value (i.e., the "value" accrued to users by utilizing forecasts in decision-making) is typically related to the forecast quality, its formulation is complex and dependent on other factors affecting the decision process (e.g., the cost assessment of action versus the losses due to missed action). Hence, quality is not equivalent to value. Nevertheless, it is possible to consider user perspectives in verification processes through the evaluation of meaningful variables and the impacts of specific thresholds, and by applying diagnostic verification approaches that examine forecast performance characteristics relevant to particular users or groups.

Forecast verification serves a number of purposes. The primary verification goals are categorized as follows:

- Scientific: to inform forecast system development and improvement;
- Administrative: for monitoring forecast performance over time; or justifying a new supercomputer acquisition;
- User-oriented: for helping users make better decisions.

Each of these purposes may require different verification approaches. Administrative users may only be interested in simple measures that are easy to compute and follow through time, whereas for scientific purposes a wider range of diagnostics is desirable to provide greater forecast performance understanding in different situations. Commonly, operational forecasting centers focus on administrative aspects, while scientists and developers focus on scientific aspects. However, incorporating information from the third aspect – forecasts users' applications – in both administrative and scientific verification efforts can often lead to more meaningful information about forecast performance.

Relevant forecast quality attributes are dependent on the type of forecast (e.g., probabilistic, deterministic) and events (e.g., categorical, continuous) of interest. Examples of forecast performance attributes include:

- *Association:* Strength of the relationship between forecasts and observations.
- *Accuracy*: Average difference (e.g. Euclidean distance) between forecasts and observations for deterministic predictions and between forecast probabilities and binary observations for probabilistic predictions.
- *Bias:* Distance between the forecast and observation average values.
- *Discrimination*: Conditioning on observed outcomes, the degree to which forecasts distinguish between different observations or events.
- *Reliability (conditional bias)*: Conditioning on the forecast, correspondence between forecast probabilities and observed relative

frequency (e.g. an event must occur on 30% of the occasions that the 30% forecast probability was issued for perfect reliability).

- *Resolution*: Conditioning on the forecasts, the degree to which observed frequency of occurrence of an event differs as the forecast probability changes.
- *Sharpness*: Degree to which forecasts deviate from the mean climatological value/category for deterministic forecasts, or from the climatological mean probabilities for probabilistic forecasts. The unconditional variation in the forecasts.

Because verification is a multi-dimensional problem, it is important to measure multiple attributes to obtain a meaningful forecast performance evaluation. That is, a single measure is unable to provide a meaningful evaluation of a forecast. Moreover, single measures can hide important information about forecast quality. For example, the root mean squared error (section 4.1) incorporates information about both bias and variance of errors; to avoid confusion about the source of a poor score it is important to consider these two features individually.

Specific forecast types may require different treatment from other forecast types, and may also create opportunities for novel evaluations. In particular, S2S forecast characteristics may lead to consideration of the S2S verification problem as somewhat different from verification at other time scales. For example, as S2S models are tuned to represent meteorological phenomena on the sub-seasonal time-scale (with a range that covers from day 15 to day 60 in some models), they are naturally suited for investigating seamless verification across the weather and seasonal time scales (Zhu et al., 2014; Wheeler et al., 2017). Another particular aspect of S2S verification is the special challenge of dealing with inhomogeneities in ensemble size between hindcasts and forecasts when evaluated together (Weigel et al. 2008), a challenge also faced in seasonal forecasting. Various studies investigated the effect of ensemble size on probabilistic forecast quality including attempts to remove the dependence on ensemble size in some verification scores to allow comparison [e.g. Richardson (2001), Muller et al. (2005), Weigel et al. 2007, Ferro (2007) and Ferro et al (2008)], being therefore relevant for S2S verification. An additional challenge for S2S verification is the need to evaluate more than one variable simultaneously for some forecast types (e.g., bivariate attributes of MJO, see section 5.3).

This chapter provides a brief overview of relevant methods for S2S forecast verification. Several additional resources exist and provide further details regarding forecast verification methods, including Wilks (2011), Jolliffe and Stephenson (2012) and a website coordinated by the WMO's Joint Working Group on Forecast Verification Research (https://www.wmo.int/pages/prog/arep/wwrp/new/jwgfvr.html). Section 2 focuses on the initial steps in the verification process: factors affecting verification studies design. Section 3 considers the issues associated with identifying appropriate observations for use in verification, and some of the issues resulting from their uncertainties. Commonly used verification measures are introduced in Section 4, and current S2S verification practices are described in Section 5. Finally, Section 6 includes a summary and recommendations.

2. Factors affecting the design of verification studies

Various factors need consideration prior to computing verification scores. The forecast type being verified is particularly important. S2S forecasts are often probabilistic rather than deterministic, and multi-category (e.g. below normal, normal and above normal). Dichotomous (yes/no) and continuous deterministic forecasts are less common, especially for user applications. This section lists the key factors and questions to be considered in the design of a verification framework or study, beginning with developing an understanding of what is required, and the target user/audience.

2.1 Target audience

The target audience is a key determinant in how a verification study should be designed, and therefore should be identified first. For example, is the verification for a model developer or for an end-user? What are the forecast performance aspects the target audience cares about? What forecast type will be evaluated? Are there user-specific thresholds to be considered? What is the scope of the verification for the user and how will the verification results be used? Should the target audience influence how verification results are presented? What complexity of metrics is appropriate? Less scientific audiences require simpler, more intuitive metrics and graphics.

2.2 Forecast type and parameters

The verification methodology is tailored to the forecast type and characteristics of the parameter to be evaluated: for example, is the forecast deterministic or probabilistic? Is it a point forecast or spatially defined? Is the variable smooth or episodic? S2S forecasts are often expressed as the likelihood of a particular weather regime, positive or negative anomalies, or multi-category (often tercile) probability forecasts. The use of anomalies is widespread and requires taking account of model climate drifts and biases. In this context it is important to identify relevant thresholds for defining the events of interest to be verified. S2S forecasts are often area-based, but can also be site-specific. The spatial and/or temporal resolution may require an analysis of representativeness (section 3), a potential issue arising when pairing gridded forecasts with observations or analyses.

2.3 Nature of available observations

Suitable and reliable observations are crucial for attaining informative verification results. It is fundamental to have observations able to capture the events the forecasts attempt to predict. What observational resolution (temporal and spatial) is required to adequately verify the forecasts? This may depend on the parameter; for example, precipitation and temperature spatial and temporal variability are very different. The impacts of inadequate (inhomogeneous) spatial and temporal sampling and observation uncertainty on verification can be large, therefore it is important to understand and take these known and unknown uncertainties into account. Questions such as the following are important to answer: Are the observations quality-controlled? Are faulty measurements corrected or disregarded? Is model information used in the quality control? How large is the observation uncertainty, and are its sources fully known? How can this uncertainty information (or lack thereof) be included in the verification results and their interpretation?

2.4 Identification of appropriate methods and metrics

Once the verification goals and purposes have been established according to user needs, and the characteristics of the available data are established, then appropriate methods and metrics can be chosen. The objective is to identify multiple verification attributes to address the questions of interest, and find graphical presentations aligned with the requirements identified in sections 2.1 and 2.2, taking account of the data issues discussed in section 2.3. Section 4 provides a summary of verification metrics used to assess the most common attributes.

3. Observational references

Observations are the cornerstone of verification. However, reliable, long-term and model-independent observations are difficult to find. This is particularly challenging for S2S where daily resolution precipitation and near surface temperature data is needed for user oriented forecasts (such as to calculate weekly averages), as opposed to monthly values, while long time series are still required. Besides, accounting for observation uncertainty in verification practices is an unresolved challenge in verification research and practice. Verification practitioners need to recognize uncertainties sources in observational datasets (e.g., measurement errors, remote-sensing retrieval algorithms, inhomogeneous and incomplete spatial and temporal sampling, time series standardization and homogenization), and their effects on verification statistics. It is also important to acknowledge model-dependencies of the verifying observation (e.g., calibration and quality control often use a model analysis as the reference) in order to correctly interpret verification results. This section reviews some of the challenges in the quest for appropriate observational references for verification purposes.

Long time series (around 30 years of measurements) are often required for climate forecast evaluation and to serve as climatological reference in S2S forecasting and verification. These time series are often affected by break-points (e.g., due to instrument replacement) and therefore complex procedures are needed to homogenize and standardize the data (e.g., Vincent and Mekis 2006). These procedures enable producing temporally coherent time series, but can affect the measured values (e.g., extremes) and introduce uncertainties in verification datasets.

Verification against point observations can suffer from representativeness issues: a point-wise measurement might differ substantially from a model value for the nearby grid cell simply because the model value is conceived to represent a grid-box average (e.g., precipitation) while the station measurement reports the value of a sub-grid phenomena (e.g., a convective cell), which is not represented by the model. Typically, due to the representativeness issue, coarse resolution models underestimate precipitation extremes; finer model spatial resolution results in better representation of intense precipitation. Similarly, coarser resolution models more often predict trace amounts than finer resolution models, leading to a positive bias for small precipitation quantities (see Figure 1).

Gridded observations are usually obtained from remote sensing instruments on satellites or from ground-based radar networks. Satellite-based products can provide gridded measurements of temperature, humidity, cloud cover, soil moisture, and sea-ice concentration and thickness. Radar-based products provide quantitative precipitation estimates. These physical variables are obtained from satellite retrieved radiances and radar backscattered reflectivities, based on remote-sensing statistical and physical assumptions (e.g., the Marshall and Palmer (1948) Z-R relationship to convert reflectivity to precipitation rate). To mitigate the effects of these assumptions (and associated uncertainties), verification can be performed with a model-to-observation approach, for example by comparing model-simulated brightness temperature directly to satellite retrieved radiances. Data assimilation algorithms are often used to harmonize, merge and quality control satellite and radar-based gridded observations, introducing model characteristics/dependence on these observations. Finally, gridding procedures (such as kriging) can introduce synthetic features, and consequently affect verification statistics.

Verification practices require that forecast and observed values are matched in space (and time). Caution is needed when choosing appropriate interpolation procedures because interpolation can alter the forecast/observed values, affecting verification statistics. For example, bilinear precipitation interpolation often introduces small trace and lower extreme values; cubic interpolation often introduces small negative precipitation values. Areal-conservative interpolation is best for precipitation upscaling (from high to low resolution grids) because model precipitation values usually represent a grid-box average, whereas a nearest-point interpolation is best for adjusting precipitation on two grids with similar resolution. Spatially smooth variables (e.g., temperature, geopotential height) are often interpolated using bi-linear or bi-cubic schemes. Neighborhood verification approaches relax the exact spatio-temporal co-location for matching forecast and observations. These approaches (as well as some spatial verification distance metrics) do not require interpolation, and therefore avoid the related issues.

Verification against a model-generated analysis is often performed because of conveniences including: i) representativeness issues, quality control and gridding is addressed by data assimilation algorithms used for analysis generation; ii) observations are spatially defined with no spatio-temporal gaps. However, a forecast model verification assessment against its own model-based analysis is affected by inter-dependence and results must be interpreted with caution. Park et al (2008) demonstrated that verification against model-based analyses strongly favors the model used to produce the analysis: for fairness in model inter-comparisons, verification against one's own analysis is therefore often adopted (at the cost of losing a single unique reference for all models). A best practice to reduce the model-analysis dependence effect could include verifying against analysis at grid points where an observation has recently been assimilated (e.g., Lemieux et al, 2015). The use of the model background state in data assimilation algorithms nudges the observations toward the model climatology, which can affect scores based on climatology. Even if not used directly for verification, data assimilation can be exploited in verification practices to provide estimates of observation uncertainties and representativeness differences.

Finally, note that all verification procedures which use model-influenced observations via the analysis (observations are nudged towards the model climatology and upscaled to the model grid) and/or quality control (filtering out of observations which differ significantly from a short-range model forecast) reduce the verification results utility for all users outside the modeling community, and generally lead to overestimating the model forecasts quality. However, it is worth noting that S2S forecast verification mostly use re-analysis/analysis as reference datasets. Another important issue for S2S verification is the inconsistency in the computation of reference anomalies used to verify the forecasts (usually computed relative to the past 20-30 years) with re-analyses and operational (real-time) analyses, which are often based on different model versions. The difference is particularly large for surface parameters and can contaminate the reference anomalies used to verify the forecasts, which are computed by subtracting the operational (real-time) analysis from the past 20 or 30 years re-analysis climatological mean, because re-analyses are usually not available in real-time and operational analyses are not available in the past.

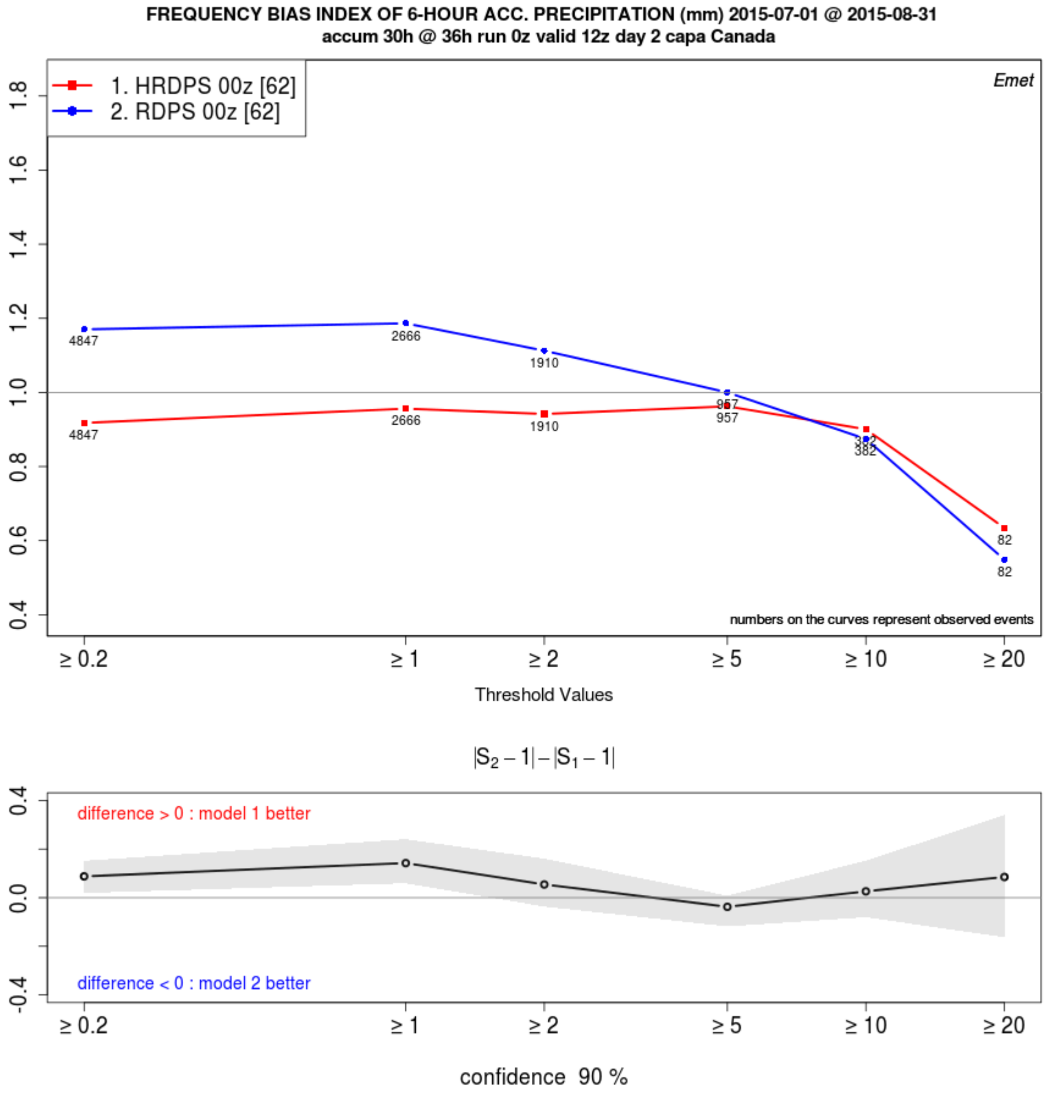


Figure 1: Frequency Bias (see section 4.2) for 6h accumulated precipitation (from 30 to 36 UTC) for the Canadian RDPS (10km resolution) and HRDPS (2.5 km resolution), for the summer 2015 against CaPA station measurements over Canada. The model with coarser resolution exhibits a larger and positive bias for smaller precipitation accumulations, and a more severe underestimation of high precipitation values, with respect to its higher resolution counterpart.

4. Review of most common verification measures

As outlined in section 2, verification metrics selected for a specific application depend on several factors; the most important are:

a. The needs of the users of the verification results;
b. The characteristics of the variable being verified;
c. The nature of the forecast and available observations.

Considering factors b and c leads to classification of forecast variables and associated metrics into the following groups:

i. Deterministic variables (forecasts and observations) – Characterized by specific variable values expressed in physical units, for example, temperature in degrees Celsius. Deterministic variables are further divided into (quasi) continuous, such as temperature, which take any physically plausible value, and categorical, which are characterized by two or more ranges of values (categories) separated by one or more predetermined thresholds. Thresholds may have physical meaning. For example, the 0.5 mm daily rain threshold is often used to separate rainfall into “no rain” and “rain” categories. Thresholds may also be set to values particularly meaningful for forecast users (e.g., setting a 50 mm threshold in 24h to indicate flooding risk). The set of categories thus defined is mutually exclusive (no overlap in values) and exhaustive (covers the whole range of possible values).

ii. Probabilistic forecasts – Forecasts indicating the probability of occurrence for pre-defined categorical variables, or the forecast probability distribution for all possible variable values. Probability forecasts are usually verified with respect to deterministic observations, even though the observations may be subject to uncertainty. If observational uncertainty estimates are available, these can be used in probabilistic forecast verification (e.g. Candille et al. 2007). Most metrics described in this section can be generalized to incorporate observational uncertainty.

iii. Spatial verification metrics. These methods are designed to account for the spatial nature of the forecast variable and corresponding observations (e.g. How does the shape of a forecast feature, such as a cloud band, compare with the observed shape?). They can be applied to deterministic or probabilistic forecasts, though the former is more common and probably carries higher physical meaning.

This section briefly describes and summarizes common metrics applicable to specific verification problems and data types. The forecast attributes (see section 1) assessed by each metric are identified.

4.1 Metrics for continuous deterministic forecasts

Table 1 summarizes the most common metrics used to verify deterministic continuous forecasts. The linear bias (B) identifies the average error for the verification sample, which is also implicitly included in both the mean absolute error (MAE) and the root mean square error (RMSE). Sometimes the bias is removed before computing the RMSE. Such removal not only reduces the RMSE, but also implies that only the variable portion of the error is assessed. However, presentation of B and the bias-corrected RMSE together disentangles these two aspects of performance, which together compose the RMSE [see Murphy (1988) for the decomposition of the mean square error (MSE)], and allows clearer understanding of the forecast errors. Comparing MAE and RMSE magnitudes for a particular sample gives an idea about the variability of the errors. The lower the variability, the smaller the difference between the two, since large errors are more heavily penalized by the RMSE. Thus the RMSE is favored when larger errors are considered relatively more important than smaller errors.

Skill scores (SS) measure forecast accuracy relative to the accuracy of a reference forecast. Most skill scores are in the general format shown in Table 1, which defines skill as the fractional improvement of the forecast score compared to the score for the reference-for-comparison. If the score for the forecast is worse than the reference, then the skill is negative. When the reference forecast accuracy is very high, and/or when the sample size is small, the skill score can become unstable, with a small denominator. For this reason, skill scores are always computed using the final summation score for a particular dataset, not for the individual cases. Skill scores commonly use the MAE, the mean square error (MSE) or the RMSE as the *score.*

While climatology (sample mean, or long-term climatology if known), random chance and persistence (the last available observation) are the most commonly used reference forecasts, sometimes skill scores are used to compare two competing forecasts, with the score for the poorer or older model version replacing the reference forecast. When the reference forecast is the climatology for the verification sample, the skill score is the same as the reduction of variance or fraction of variance explained, which is the same as the square of the correlation coefficient between the forecasts and observations. This interpretation is more complicated when the reference is a second forecast.

The Pearson product moment correlation coefficient (r) is often used to measure the strength of the linear relationship between forecasts and observations (the association attribute). Perfect association (i.e., r=1) is obtained when the forecasts and observations oscillate exactly in the same direction. This measure, however, only provides an indication of potential skill because correlation is insensitive to forecast biases as well as differences in forecast versus observation variances. Several of the continuous measures can be displayed simultaneously using a Taylor diagram (Taylor 2001). In particular, this diagram displays the correlation coefficient, root-mean-square difference, and the ratio of the standard deviations of the forecast and observed patterns.

Table 1: Common metrics (or scores) for verifying continuous deterministic forecasts (F_i) against the observations (O_i). Subscripts *i* refer to the i^{th} case of the verification sample; the sample of forecast and observation pairs is of size *N*; the overbar indicates sample averaging. S_f refers (usually) to either the MAE or RMSE scores computed for the N pairs of F_i and O_i according to the equations in the table; S_r refers to the same score computed using a unskilled reference forecast such as the variable mean (climatology) or the latest observed value of the variable (persistence); S_p refers to the score for the perfect forecast. For perfect forecasts where F_i=O_i for all N pairs both MAE=0 and RMSE=0 (i.e., S_p=0).

METRIC	EQUATION	ATTRIBUTE MEASURED	CHARACTERISTICS
Bias (Linear bias, B)	$B = \frac{1}{N}\left[\sum_{i=1}^{N}(F_i - O_i)\right]$	Accuracy (average error)	Estimates the persistent or average error, based on specific dataset; negative orientation (best when B=0)
Mean Absolute Error (MAE)	$MAE = \frac{1}{N}\left[\sum_{i=1}^{N}\mid F_i - O_i \mid\right]$	Accuracy	Average error magnitude, negative orientation (best when MAE=0)
Root Mean Square Error (RMSE)	$RMSE = \left[\frac{1}{N}\sum_{i=1}^{N}(F_i - O_i)^2\right]^{1/2}$	Accuracy	Average error magnitude weighted to larger errors; Negative orientation (best when RMSE=0)
Skill score (SS)	$SS = \frac{S_f - S_r}{S_p - S_r} = \frac{S_r - S_f}{S_r} = 1 - \frac{S_f}{S_r}$	Skill (general format) For negatively oriented scores, perfect score S_p=0)	Fractional improvement of the forecast over an unskilled reference. Range: -∞ to 1.
Pearson correlation coefficient (r)	$r = \frac{\sum_{i=1}^{N}(F_i - \overline{F})(O_i - \overline{O})}{\sqrt{\sum_{i=1}^{N}(F_i - \overline{F})^2}\sqrt{\sum_{i=1}^{N}(O_i - \overline{O})^2}}$	Association	Strength of the linear relationship between forecasts and observations Range: -1 to 1.

4.2 Verification methods for categorical deterministic forecasts

Categorical deterministic forecasts are often synthesized using contingency tables. Table 2 shows the contingency table and scores for a 2 X 2 (2 category) variable, all of which are functions of the four table entries: hits (*a*), misses (*c*), false alarms (*b*), and correct negatives (*d*). Fundamental score characteristics are indicated in the table. Analogous table forms and scores exist for more than two categories, but multi-categories of a single variable are often treated as sequences of 2-category problems, with boundaries at each of the thresholds in turn. Murphy and Winkler (1987) relate the contingency table and its entries to the forecast and observation joint probabilities, providing the statistical framework to interpret categorical scores as functions of joint, conditional and marginal probabilities.

Several relationships exist between the categorical scores listed in Table 2, so that a subset of those is often calculated, such as the frequency bias (FB) and the equitable threat score (ETS) or the Heidke skill score (HSS, to assess bias and accuracy/skill). The FB is not a verification score in the strict sense because it does not depend on matched forecasts and observations pairs. As a ratio of the forecast frequency to the observed frequency of each event category, it describes the forecast strategy, “overforecasting” if greater than 1 and “underforecasting” if less than 1. Several categorical scores can be displayed simultaneously in a performance diagram (Roebber, 2008).

Often the occurrence of one of the two categories is larger than the other, particularly in the case of extreme events, which are usually much less common than the corresponding “non-event”. The extremal dependence index (EDI) and the symmetric extremal dependence index (SEDI) are specifically designed to score categories with low observed event frequency (a+c)/N (called the base rate or climatological frequency). Under these conditions, scores such as the threat score (TS), the hit rate (H), the false alarm ratio (FAR), the false alarm rate (F), the ETS and the Hanssen-Kuipers discriminant score (KSS) tend artificially towards their limit values (0 or 1), rendering the interpretation of the verification results challenging. The HSS may also become unstable for low base rates because the unskilled forecast accuracy is high.

When both categories are of similar interest, H, FAR, and TS can be computed for both categories (event and non-event) separately. For example H for the non-event category is d/(b+d) and TS is d/(b+c+d). In this situation the proportion correct PC=(a+d)/N is also an informative score. PC is not recommended otherwise, because it becomes misleading when one category occurs more frequently than the other. See literature on the “Finley affair” (Murphy 1996).

H and F refer to stratifying the verification dataset in terms of (conditioned on) the observations. These scores are often used in pairs, and along with the relative operating characteristic curve (ROC) and area (ROCA) [see section 4.3] are useful for evaluating the *a posteriori* forecast quality as a basis for users’ decision-making.

FAR is different from F (and sometimes confused with it) – it is the proportion of forecasts which are false alarms, that is, it is conditioned on the forecasts. FAR is widely used, and can be controlled by the forecaster, by for example forecasting the event less often to reduce the number of false alarms. This strategy would also increase the number of missed events (*c*); therefore FAR should be used in combination with H.

The correct negatives (*d*) are sometimes hard to determine for a contingency table, since the “non-event” may be spatially and/or temporally unbounded; *d* can also be very large in the case of extreme events and overwhelm the contingency table computations. The scores H, FAR, and TS use only the other three entries of the contingency table, and therefore can be computed without estimating or taking into account correct negatives. However, all skill scores and scores useful for discrimination and decision making need correct negatives estimates. Some ways of estimating *d* for severe weather non-events are suggested in Wilson (2014), and Wilson and Giles (2013).

Table 2. Contingency table format and associated scores. The letters a, b, c and d refer to total counts of cases with the corresponding pairing of forecast and observation. Sample size is N.

<table>
<tr><th>Measure</th><th>Equation/Format</th><th>Range - Orientation</th><th>Characteristics</th></tr>
<tr><td>Contingency table</td><td>
<table>
<tr><td></td><td></td><td colspan="2">Observed</td><td></td></tr>
<tr><td></td><td></td><td>Yes</td><td>No</td><td>Total fcst</td></tr>
<tr><td rowspan="2">Forecast</td><td>Yes</td><td>a
(Hits)</td><td>b
(False alarms)</td><td>a+b
(total events forecast)</td></tr>
<tr><td>No</td><td>c
(Missed events)</td><td>d
(Correct negatives)</td><td>c+d
(total non-events forecast)</td></tr>
<tr><td></td><td>Total obs</td><td>a+c
(total events obs)</td><td>b+d
(total non-events obs)</td><td>N=a+b+c+d
(sample size)</td></tr>
</table>
</td><td>Normally as shown, columns are conditional observation totals, rows are conditional forecast totals</td><td>Equivalent to a scatter plot for categorized variable. Two by two table most common – two categories, one threshold.</td></tr>
<tr><td>Frequency Bias (FB)</td><td>$\mathrm{FB} = \frac{a+b}{a+c}; \frac{c+d}{b+d}$
Ratio between the total number of events forecast (or not forecast) and the total number of events observed (or not observed)</td><td>0 to ∞</td><td>Best score = 1. Simple comparison of forecast frequency to observed frequency.</td></tr>
<tr><td>Hit rate (H) (Probability of detection)</td><td>$\mathrm{H} = \frac{a}{a+c}$</td><td>0 to 1</td><td>Best = 1. Incomplete score – does not account for false alarms</td></tr>
<tr><td>False alarm rate (F) (probability of false detection)</td><td>$\mathrm{F} = \frac{b}{b+d}$</td><td>1 to 0</td><td>Best = 0. Can be improved by forecasting the event less often to reduce false alarms</td></tr>
<tr><td>False alarm ratio (FAR)</td><td>$\mathrm{FAR} = \frac{b}{a+b}$</td><td>1 to 0</td><td>Best = 0. Sensitive to false alarms but ignores misses. Use with <i>H</i></td></tr>
<tr><td>Threat score (TS) (Critical success index)</td><td>$\mathrm{TS} = \frac{a}{a+b+c}$</td><td>0 to 1</td><td>Best = 1. Sensitive to both false alarms and misses; ignores correct negatives</td></tr>
</table>

Equitable threat score (Gilbert skill score) (ETS)	$\mathrm{ETS} = \frac{a - a_r}{a + b + c - a_r}$ Where $a_r = \frac{(a + b)(a + c)}{N}$	-1/3 to 1; 0 indicates no skill over chance	Best = 1. TS adjusted for the number correct by chance (guessing), a form of skill score. Always < TS
Hanssen-Kuipers discriminant (KSS) (also true skill statistic TSS or Pierce skill score, PSS)	$\mathrm{KSS} = \frac{a}{a + c} - \frac{b}{b + d} = H - F$	-1 to 1; 0 indicates no discriminant ability	Best = 1. Related to the ROC area and EDI/SEDI scores. Indicates the ability of the forecast to discriminate between events and non-events, as a basis for decision-making
Heidke skill score (HSS)	$\mathrm{HSS} = \frac{(a + d) - E_r}{N - E_r}$ Where $E_r = \frac{1}{N}[(a + c)(a + b) + (c + d)(b + d)]$	-∞ to 1	Best = 1. Skill score in the general format with "chance" as the reference forecast.
Extremal dependence index (EDI)	$\mathrm{EDI} = \frac{\ln F - \ln H}{\ln F + \ln H}$	-1 to 1; 0 indicates no accuracy	Best = 1. Designed to avoid convergence to 0 or 1 for low frequency (rare) events. Most often used for verifying extreme event forecasts
Symmetric extremal dependence index (SEDI)	$\mathrm{SEDI} = \frac{\ln F - \ln H + \ln(1 - H) - \ln(1 - F)}{\ln F + \ln H + \ln(1 - H) + \ln(1 - F)}$	-1 to 1; 0 indicates no accuracy	Best = 1. Similar to EDI, but approaches 1 only for unbiased forecasts.

4.3 Verification measures for probability forecasts

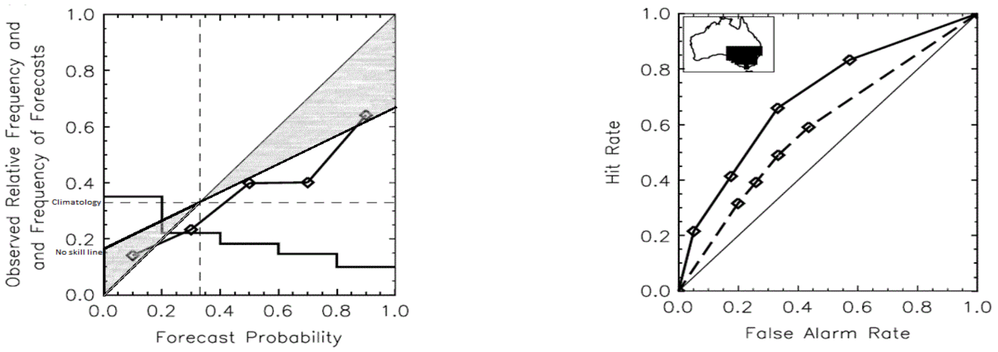
Probability forecasts are estimates of the likelihood of occurrence of an "event", which is usually defined as a category of a variable (e.g., the probability for the daily average temperature to be in the upper tercile of the temperature climatological distribution for a specific location or area). Categories are defined by thresholds as for categorical variables. Probability forecasts are difficult to verify meaningfully as single forecasts because the observation is usually treated as categorical (the event either occurred as forecast or not). Probability forecast verification thus proceeds after a sufficiently large sample of matched forecasts and observations is collected, allowing comparative assessment of the actual event occurrence frequencies with the forecast probabilities.

While probability forecasts are most often obtained from ensembles, it is worth noting that ensemble forecasts require post-processing to generate probabilities, by calculating probabilities of occurrence of events simply from the proportion of the ensemble members satisfying the threshold for the event, or by using the ensemble to estimate a full predicted distribution. Ensembles are collections of deterministic forecasts obtained from perturbed initial conditions and/or variations in the model formulation. Raw ensembles are assumed to be a random selection from the unknown conditional probability density function (pdf) of possible forecast values and/or the associated cumulative distribution function (cdf). The resulting forecast values distribution is inherently discrete given the relatively small ensemble sizes, but processing methods are available to estimate continuous pdfs. The verification methods summarized below are suitable for probability forecasts of specific events or for forecast pdfs.

Table 3. Common scores for probabilistic forecasts verification. The variables p_i and o_i refer to the ith forecast probability and ith observation in a sample of size N. The observation o_i is 0 (1) if the category predicted with probability p_i doesn't (does) occur. Subscript k refers to the kth category of a total of M categories, and $P_f(x)$ and $P_o(x)$ are the predicted and observed cdfs respectively, the latter taking the form of a step (heaviside) function with the step at the observed value of the variable x. S_f, S_p and S_r are defined exactly in the same way as Table 1.

Measure	Equation	Range - orientation	Characteristics
Brier Score (BS)	$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2$	1 to 0	Best = 0, negatively oriented, Mean squared error of the probability forecasts
Ranked Probability Score (for discrete categories)	$\mathrm{RPS} = \frac{1}{M-1}\sum_{m=1}^{M}\left[\left(\sum_{k=1}^{m} p_k\right) - \left(\sum_{k=1}^{m} o_k\right)\right]^2$	1 to 0	Best = 0, equals BS for 2 categories, for > 2 categories, sensitive to distance between forecast and observed category
Continuous rank probability score (CRPS)	$\mathrm{CRPS} = \int_{-\infty}^{\infty}\left[P_f(x) - P_o(x)\right]^2 dx$	0 to ∞	Best = 0, Compares cdf for forecast with cdf of observation. Obs cdf is step function if deterministic; result is in the units of the variable, reduces to MAE for deterministic forecast
Brier skill score (BSS), rank probability skill score (RPSS) and Continuous rank probability skill score (CRPSS)	$\mathrm{SS} = \frac{S_f - S_r}{S_p - S_r} = \frac{S_r - S_f}{S_r} = 1 - \frac{S_f}{S_r}$	-∞ to 1	Best = 1. Skill scores in the standard format for negatively oriented scores. Caution: The reference forecast is defined by the sample over which the skill score is computed. (See Hamill and Juras, 2006)

Table 3 summarizes commonly used verification scores for probability forecasts. The Brier score (BS) is generally used for probability forecasts for a dichotomous (binary) variable, while the discrete ranked probability score (RPS) is preferred when there are more than two categories. The continuous ranked probability score (CRPS) is used to evaluate the full continuous or quasi-continuous forecast cdf. The BS, RPS and CRPS all measure the attribute *accuracy*, while the corresponding scores shown on the bottom row of the table measure *skill*. The three scores can be partitioned into three components representing the attributes *reliability*, *resolution* and *uncertainty,* the latter being a function of the observations only. The reliability (or attributes) diagram and the ROC curve offer a concise and convenient graphical representation of most probability forecasts attributes listed in section 1.



(a) (b)

Figure 2: Attributes/reliability diagram (a) and ROC plot (b) for the event average week 3-4 (second fortnight) maximum temperature forecasts in the upper tercile for Southeast Australia for spring start months. See text for description and interpretation. (Adapted from Figure 7 of Hudson et al 2011)

Figure 2a shows an example of an attributes diagram (which is based on a reliability diagram) produced by first binning the verification sample according to forecast probability, and next computing the observed event frequency for all of the forecasts in each bin. The diagram is a plot of the observed frequency versus the forecast probability for each bin. Five bins are used (0 to 20%, 20 to 40%, 40 to 60%, 60 to 80%, 80 to 100%). The points are plotted at the mid points of the five bins, but it is possible and perhaps more accurate to plot the points at the actual mean forecast probability for each bin. Forecasts are considered perfectly reliable if the points lie along the 45^o diagonal, indicating that, on average, the forecast probability is equal to the observed event frequency. Dashed lines are drawn horizontally at the climatological event frequency (the base rate), and vertically at the average forecast probability. Comparing these two lines indicates whether the event is, on average, overforecast or underforecast. In the example, there is no (unconditional) bias. The average forecast probability is about 33%, equal to the observed frequency.

The line which bisects the angle between the diagonal and the climatology line is known as the "no skill line". On this line, the resolution of the forecasts is equal to the reliability component (and of opposite sign), so that skill with respect to sample climatology along this line, as computed by the BSS, is zero. When the plotted curve lies within the shaded area, the forecasts have skill with respect to climatology. In the example there is apparently little skill, but there is an indication of resolution in the forecasts since the plotted line is inclined with respect to the base rate line. The plotted curve presents a shallower than 45^o angle, indicating that the forecasts are overconfident. The highest probabilities are overforecast while the lowest probabilities are slightly underforecast. Curves lying at a steeper than 45^o angle (in the area between the diagonal and the vertical dashed line) may be skillful, but are said to be over-resolved and/or underconfident. This situation does not often happen in practice. Finally, the histogram on the reliability diagram represents the percentage of the forecast probabilities sample falling into each bin, known as the “sharpness diagram”. Sharp forecasts have *u*-shaped histograms presenting high frequencies for near 0 and 100% forecast probabilities. In the example, forecasts are not particularly sharp, and skewed towards the lowest probabilities, with probabilities of less than 20% forecast nearly 40% of the time.

The ROC curve (Figure 2b) and the area under the curve (ROCA) measures the ability of the forecast to discriminate situations leading to events of impact from those which do not. Such discriminating ability is useful for decision-makers in deciding whether to take action to minimize adverse weather/climate impacts. The curve is obtained as follows:

1. Organize the verification sample in ascending order of the predictive variable (usually forecast probabilities, but can also be a physical variable such as precipitation amount). If the forecasts are from ensembles, each set of ensemble forecasts can be ordered and pooled over all verification sample cases. The associated observation is binary (1 or 0) according to whether the event occurred or not for each case. For an ensemble forecast the observation value 0 or 1 is assigned to all members of that particular ensemble.
2. For each unique prediction value, considered as a prediction threshold for the event of interest, compute the false alarm rate and hit rate for the resulting contingency table.
3. Plot the hit rate against the false alarm rate. The result is a stepwise graph, approximating a curve. The more points that are possible (the more unique forecast values exist in the dataset), the “smoother” the curve will be.
4. The area under the curve can be computed by triangulation, using all plotted points.

Examples of the computation of the ROC by this method are shown in Mason and Graham (2002). It is common practice to bin the data into forecast categories, often forecast probability deciles, or as in figure 2b, in pentiles. This results in fewer points to estimate the curve, possibly leading to an underestimation of the ROCA. For example, if the frequency of occurrence of the event for a 5% forecast is lower than for 15% forecasts, then this discrimination information is lost if the data are binned into 20% bins. However, the tradeoff is that there must be enough cases in each bin to support the plotted points, or the plotted curve will be noisy and confidence in the location of the points will be low. Underestimates of the ROCA can be avoided by fitting a binormal model to binned data (Wilson 2000). The binormal model is described in Mason (1982).

Discrimination ability is indicated if the ROC curve lies in the upper left half of the diagram. The closer the curve lies to the upper left corner, the better the forecast discrimination ability. The diagonal is the “no skill”, or “no discrimination” line. This means that the forecast probabilities distribution when the event occurs is no different from the forecast probabilities distribution when the event does not occur, and therefore the user has no basis to decide whether to take action or not. The ROCA is the total area between the lower right corner and the ROC curve. ROCA larger than 0.5 indicates discrimination ability. Perfect discrimination (ROCA=1) occurs when there is no overlap at all between the conditional forecast distribution when the event occurred and the conditional forecast distribution when the event did not occur. Sometimes the ROC score is expressed as $2\mathrm{ROCA}-1$ to give a positively oriented score varying between 0 and 1.

Figure 2b shows two ROC curves. The solid line is for S2S model probability forecasts for the event “second fortnight averaged maximum temperature in the upper tercile for south-eastern Australia”. The dashed line is obtained by assuming the first fortnight average temperatures persist for the second fortnight. The plot indicates modest discrimination, with some improvement of the S2S model (ROCA=0.70) over the persistence forecast (ROCA=0.59).

One cautionary note is needed with regard to application of ROC plots and skill scores with respect to climatology. The relevant standard of comparison is always the mean observation of the sample used to compute the score. For the ROC, discrimination of all variation sources from the overall sample mean is credited. This effect is discussed in Hamill and Juras (2006).

4.4 Spatial Methods

Meteorological variables defined over spatial fields are characterized by spatial structures and features. Traditional point-by-point verification approaches do not account for the intrinsic spatial correlation existing between nearby grid-points. This practice leads to double penalties (associated with small spatial displacements) and limited diagnostic power (traditional scores do not inform on displacements or error scale dependence). To address these issues, several spatial approaches have been developed and applied to weather forecasts in the past two decades. Spatial verification techniques aim to:

i) account for spatial structure and features;
ii) provide information on the forecast error in physical terms (e.g., diagnose the location error as distances in km); and
iii) account for small time-space uncertainties.

Spatial verification approaches are categorized in five classes:

1. *Scale-separation* approaches involve decomposition of the forecast and observation fields into scale components using a single band spatial filter (Fourier transforms, wavelets, spherical harmonics), followed by a traditional verification on each spatial scale component. The rationale is to provide information on physical processes associated with weather phenomena on different scales (frontal systems versus convective precipitation; planetary, synoptic and sub-synoptic scales). These approaches enable assessment of bias, error and skill on each individual scale; are used to analyze predictability scale-dependence (by determining the no-skill to skill transition scale); and to assess the forecast versus observation scale structure. Scale-separation techniques have been successfully applied both to weather and climate studies (e.g.

Casati 2010; Jung and Leutbecher 2008; Denis et al. 2002, 2003; Livina et al. 2008) and can be useful in the S2S framework.

2. *Neighborhood* methods (Ebert 2008) relax the requirement for an exact observation-forecast location match, and define a neighborhood (both in space and time) where the forecast and observation are matched. Data treatment within the neighborhood differentiates the verification strategies which include simple averaging (equivalent to upscaling, Yates et al, 2006); comparison of forecast versus observed event frequencies (Roberts and Lean; 2008); evaluation of different attributes of the forecast versus observed pdf (Marsigli et al, 2005); application of probabilistic and ensemble verification approaches to assess the forecast pdf within the observed neighborhood (Theis et al, 2005). Neighborhood approaches are suitable for comparing higher versus coarser resolution models. Moreover, they enable probabilistic evaluation of deterministic forecasts.
3. *Field deformation* techniques use a vector field to deform the forecast field towards the observed field until an optimal fit is found (by maximizing a likelihood function). A scalar (amplitude) field is then applied, in order to correct the intensities of the deformed forecast field to those of the observed field. These morphing techniques were originally developed for data assimilation and nowcasting (Nehrkorn et al. 2003; Germann and Zawadzki, 2004), and have only recently been used in verification (Keil and Craig 2007, 2009; Marzban and Sandgathe 2010; Gilleland et al. 2010).
4. *Feature-based* verification techniques (Ebert and McBride 2000; Davis et al. 2006a,b) first identify and isolate features in forecast and observation fields (by thresholding, image processing, using composites, cluster analysis), and then assess different attributes (displacement, timing, extent, intensity) for each pair of observed and forecast features.
5. *Distance measures* for binary images assess the distance between forecast and observation fields by evaluating the (geographical) distances between all the grid-points exceeding a selected threshold. These metrics were developed in image processing for edge detection and/or pattern recognition (Dubuisson and Jain 1994; Baddeley 1992a,b) and only recently used for verification purposes (Schwedler and Baldwin 2011; Gilleland, 2011; Dukhovskoy et al. 2015). The distance measures are sensitive to differences in object shape and extent in addition to the distance/displacement between forecast and observed features, and thus are considered a hybrid between field-deformation and feature-based techniques.

5. Types of S2S forecasts and current verification practices

5.1 Deterministic S2S forecast verification practices

Sub-seasonal forecasts are often presented as weekly averages for the forthcoming four weeks, either defined as averages over days 1 to 7 (week 1), 8 to 14 (week 2), 15 to 21 (week 3), and 22 to 28 (week 4) as in Li and Robertson (2015), or as averages over days 5 to 11 (week 1), 12 to 18 (week 2), 19 to 25 (week 3), and 26 to 32 (week 4) as in Weigel et al. (2008). Some studies [Hudson et al. (2011, 2013)] investigate averages over days 1 to 14 (first fortnight) and 15 to 28 (second fortnight).

As in weather and seasonal forecasting practice, the mean of the available ensemble members is commonly used as an estimate of the forecast distribution central value. Deterministic forecasts are expressed as ensemble mean anomalies, computed by subtracting the ensemble mean forecast from the model long term mean (climatology) estimated using retrospective forecasts produced for a number of previous years for a first-order model bias correction. This procedure used for computing ensemble mean anomalies is typically lead time dependent.

The simplest S2S verification practice is eyeball (visual) comparison of the forecast ensemble mean anomaly with the corresponding observed anomaly. As shown in Figure 1 of Vitart et al. (2017) one can, for example, visually compare 2-meter temperature ensemble mean forecast anomaly maps for different models with the observed anomaly. Eyeball comparison is useful for an initial qualitative assessment of specific forecasts but is prone to subjective interpretation biases, and therefore must be used with caution. Quantitative assessment obtained by computing verification metrics based on a collection of past forecasts and observations provides a more complete view of forecast quality.

A common verification practice for deterministic (ensemble mean) S2S forecasts is to compute the linear correlation between the forecast and observed anomalies at each grid point (over the available retrospective forecasts) and produce a map with the obtained values [see Figures 1 of Hudson et al. (2011) and Li and Robertson (2015), and Figure 10 of Weigel at al. (2008)]. The Pearson product moment correlation coefficient is often used for this purpose, providing an association measure (see section 4.1). However, due to its insensitivity to forecast biases, complementary accuracy metrics are required to quantify forecast errors. A standard S2S metric used for this purpose is the linear bias (see section 4.1). Figure 3 of Weigel et al. (2008) provides an example of 2-meter temperature ensemble mean bias for weeks 1 to 4 retrospective forecasts.

Deterministic S2S forecast skill can be estimated using the mean squared error skill score (MSSS) as performed by Li and Robertson (2015). The MSSS is based on the MSE, an accuracy measure similar to the RMSE (see section 4.1) with the main difference being that the square root needed for the RMSE is not computed for the MSE. In the examples shown in Figures 13 to 15 of Li and

Robertson (2015) the reference set of forecasts used to compute the MSSS were climatological forecasts given by the climatological average rainfall for a given weekly average. The maximum MSSS equals unity and is obtained for perfect forecasts with null MSE. Negative values indicate that the forecasts are less accurate than the reference climatological forecast.

5.2 Probabilistic S2S forecast verification practices

A common procedure in S2S probabilistic forecast verification is to construct ROC plots and reliability diagrams as described in Section 4.3, or to compute the RPSS and construct reliability diagrams as in Vigaud et al. (2017a,b). Figure 12 of Vitart and Monteni (2010) and Figure 3 of Hudson et al. (2011) show additional ROC plots and reliability diagrams S2S examples for a collection of forecasts aggregated over a number of grid points within a pre-defined area/region. The area under the ROC plot provides an indication of the ability of the forecasting system in successfully discriminating occurrence from non-occurrence of the event of interest (i.e., how forecast probabilities vary when stratified on the observations). The reliability diagram provides a graphical interpretation of probabilistic forecast quality in terms of reliability (how well forecast probabilities match the observed frequency of the event of interest) and resolution (how the observed frequency varies when the data are stratified by the forecast probabilities).

By computing the ROC area at each grid point and mapping the collection of obtained values one can have a spatial idea of forecast discrimination ability, particularly for regions exhibiting ROC area above 0.5 [the reference value for unskillful forecasts with equal (50%) probability of distinguishing/discriminating events from non-events]. Figure 2 of Hudson et al. (2011) shows examples of ROC area maps for probabilistic forecasts of precipitation averaged over the first and second fortnight for events defined as precipitation in the lower and upper terciles.

5.3 Madden and Julian Oscillation (MJO) forecast verification

A specific type of sub-seasonal forecast is the forecast of the Madden and Julian Oscillation (MJO, Madden and Julian 1971, 1972, 1994; Zhang 2005), which usually emerges as enhanced convection over the Tropical Indian Ocean and propagates eastward along the Equator. MJO forecasts and the associated verification are important to both model developers and forecasters in order to provide information about model behavior and performance in representing tropical precipitation.

MJO forecasts are distinct from traditional weather and climate forecasts because they are displayed in a two dimensional phase space represented by two so-called Real-time Multivariate MJO indices (RMM1 and RMM2) as defined by Wheeler and Hendon (2004). Figure 3a shows an example of an MJO forecast initialized on 1 January 1986 for the following 41 days. The initial points for the observations (blue line) and ensemble mean forecasts (red dashed line) are indicated with large brown dots. The small black dots are separated by 5 days. Counter-clockwise progression indicates eastward MJO signal propagation. The MJO strength is measured by the distance of each point in the phase space diagram to the origin. The central circle represents one standard deviation and is usually considered the threshold for defining an active MJO signal. The observed RMM1 and RMM2 are the principal component time series of the first and second leading modes of the combined empirical orthogonal function (EOF) analysis of daily outgoing longwave radiation (OLR), 850 hPa and 200 hPa zonal wind anomalies latitudinally averaged from 15^{o}S to 15^{o}N. Both RMM1 and RMM2 are normalized by the observational standard deviation, resulting in indices with zero mean and unit variance. See Rashid et al. (2011) and Gottschalck et al. (2010) for additional information on how the forecast RMM1 and RMM2 are computed.

Figure 3b shows the schematic for a pair of points [O(t) and F(t,τ)] in the two dimensional MJO phase space represented by the RMM1 and RMM2 indices (horizontal and vertical axis, respectively). The point O(t) highlighted with a blue dot represents the location of the observed MJO signal at time t. The point F(t,τ) highlighted with a red dot represents the forecast MJO signal at time t produced τ days earlier. The points O(t) and F(t,τ) in Figure 3b illustrate, for example, the fourth black dots after the initial large brown dots shown in Figure 3a representing a forecast for time t equal to the 20 January 1986 produced in the previous 1st January 1986 (τ=20 days lead forecast). The blue solid line connecting the origin of the phase space plot to the point O(t) graphically illustrates the observed MJO signal. The projections of this signal along the horizontal and vertical axes are illustrated in Figure 3b as $a_1(t)$ and $a_2(t)$ and represent the observed RMM1 and RMM2, respectively. The red solid line connecting the origin of the phase space plot to the point F(t,τ) graphically illustrates the forecast MJO signal for time t produced τ days earlier. The projection of this signal along the horizontal and vertical axis is illustrated in Figure 3b as $b_1(t,τ)$ and $b_2(t,τ)$ and represents the τ days lead RMM1 and RMM2 forecasts for time t, respectively.

As the RMM1 and RMM2 axis in Figure 3b are orthogonal, the observed a(t) and forecast b(t,τ) MJO amplitudes are expressed as

$$a(t) = [a_1(t)^2 + a_2(t)^2]^{1/2} \qquad (1)$$

$$b(t,\tau) = [b_1(t,\tau)^2 + b_2(t,\tau)^2]^{1/2} \qquad (2)$$

and the observed (φ) and forecast (Θ) MJO phases, represented by the angles between the blue and red lines and the horizontal RMM1 axis, respectively, are expressed as

$$\phi(t) = \tan^{-1}\left(\frac{a_2(t)}{a_1(t)}\right) \qquad (3)$$

$$\theta(t,\tau) = \tan^{-1}\left(\frac{b_2(t,\tau)}{b_1(t,\tau)}\right) \qquad (4)$$

Following Rashid et al. (2011) the amplitude A(τ) and phase P(τ) errors for a collection of N forecast and observed MJO pairs as function of forecast lead time τ are defined as

$$A(\tau) = \frac{1}{N}\sum_{t=1}^{N}[b(t,\tau) - a(t)] \qquad (5)$$

$$P(\tau) = \frac{1}{N}\sum_{t=1}^{N}\tan^{-1}\left(\frac{a_1 b_2 - a_2 b_1}{a_1 b_1 + a_2 b_2}\right) \qquad (6)$$

The amplitude error verification metric A(τ) is similar to the linear bias (see section 4.1). Both A(τ) and P(τ) measure accuracy in term of the average error. A(τ) is negatively oriented (best forecasts have A(τ)=0). P(τ) expresses the mean angle difference (Θ-φ) of the forecast Θ and observed φ MJO phases over the N available pairs. P(τ) is positive if the forecast phase on average leads the observed phase. Note that to obtain Eq. (6) one needs to use cross and dot product properties in the process of finding the angle (Θ-φ) between the observed (blue line) and forecast (red line) MJO signal.

Lin et al. (2008) introduced the following metrics for evaluating the quality of the bivariate MJO forecasts displayed in the RMM1 versus RMM2 phase space: the bivariate correlation r(τ), the root mean square error RMSE(τ) and the mean square skill score MSSS(τ) following Murphy (1988).

$$r(\tau) = \frac{\sum_{t=1}^{N}[a_1(t)b_1(t,\tau) + a_2(t)b_2(t,\tau)]}{\left(\sum_{t=1}^{N}[a_1(t)^2 + a_2(t)^2]\right)^{1/2}\left(\sum_{t=1}^{N}[b_1(t,\tau)^2 + b_2(t,\tau)^2]\right)^{1/2}} \qquad (7)$$

$$\mathrm{RMSE}(\tau) = \left(\frac{1}{N}\sum_{t=1}^{N}\varepsilon(t,\tau)^2\right)^{1/2} \qquad (8)$$

where

$$\varepsilon(t,\tau)^2 = [a_1(t) - b_1(t,\tau)]^2 + [a_2(t) - b_2(t,\tau)]^2 \qquad (9)$$

See figure 3b for a graphical representation of ε(t,τ).

Finally,

$$\mathrm{MSSS}(\tau) = 1 - \frac{\mathrm{MSE}(\tau)}{\mathrm{MSE}_C} \qquad (10)$$

where

$$\mathrm{MSE}(\tau) = \frac{1}{N}\sum_{t=1}^{N}[a_1(t) - b_1(t,\tau)]^2 + [a_2(t) - b_2(t,\tau)]^2 \qquad (11)$$

and

$$\mathrm{MSE}_C = \sum_{t=1}^{N}[a_1(t)^2 + a_2(t)^2] \qquad (12)$$

is the mean squared error for the climatological (unskillful) forecast that always issues an absent MJO signal RMM1=RMM2=0 for all t and τ [$b_1(t,\tau)=b_2(t,\tau)=0$] and is equivalent to the observed (climatological) variance of the MJO.

The bivariate correlation r(τ) is an association measure examining the strength of agreement (or disagreement) between the observed (φ) and forecast (Θ)

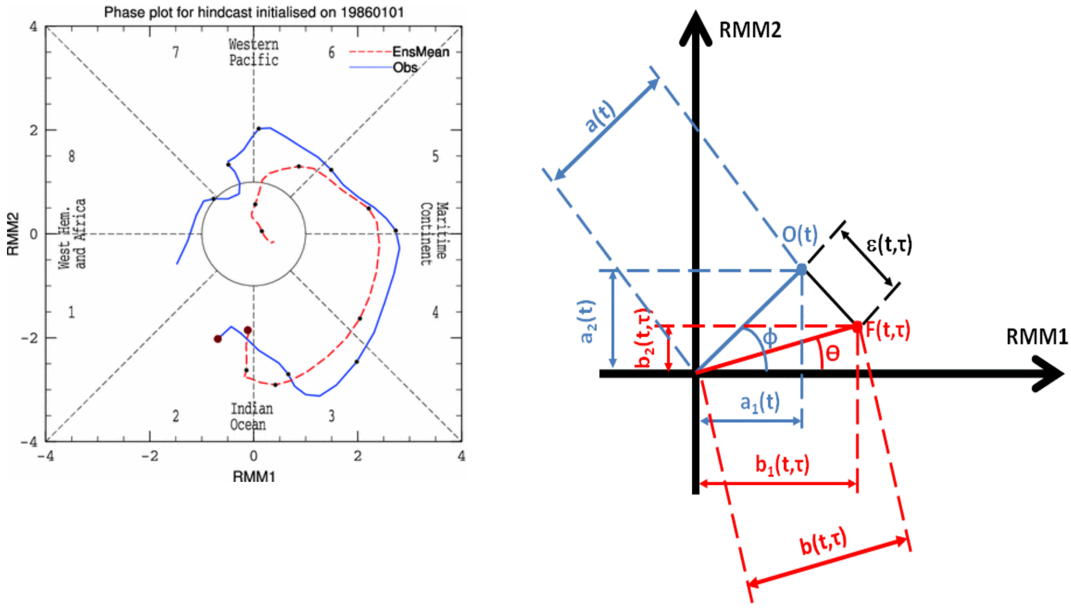
MJO phases, but is insensitive to MJO amplitude errors (biases in the magnitude of the forecast MJO signal). The RMSE(τ) is a simultaneous accuracy measure of both phase and amplitude of the MJO similar to the RMSE earlier introduced in section 4.1.

The upper limit for the bivariate correlation r(τ) is obtained for perfect forecasts indicating an exact match between the forecast and observed phases of the MJO (when $\Theta=\varphi$) and equals unity. The lower limit for r(τ) is obtained for forecasts indicating an opposite match between the forecast and observed phases and is equal to −1 (when $\Theta=\varphi+180^o$). For perfect forecasts with $a_1(t)=b_1(t,\tau)$ and $a_2(t)=b_2(t,\tau)$ the bivariate RMSE(τ) equals 0. For the climatological forecasts [in the absence of an MJO signal and thus $b_1(t,\tau)=b_2(t,\tau)=0$], the bivariate RMSE(τ) equals $2^{1/2}$ because the variance of each of the two observed RMM indices [$a_1(t)$ and $a_2(t)$] is equal to 1. Forecasts are generally considered skillful if their RMSE(τ) is less than $2^{1/2}$ (the RMSE(τ) for climatological MJO forecasts). For forecasts with observed amplitude but completely random phase [a persistence forecast at very long lead time such that $a_1(t)=b_1(t,\tau)$ and $a_2(t)=-b_2(t,\tau)$] the RMSE(τ) asymptotes to 2.

The mean squared skill score MSSS(τ) provides a relative measure of skill for the MJO forecasts compared to the climatological forecast that indicates an absent MJO signal [$b_1(t,\tau)=b_2(t,\tau)=0$]. Perfect forecasts with MSE(τ)=0 have MSSS(τ)=1. Forecasts with errors as large as the climatological variance [$MSE(\tau)=MSE_C$] have a null skill score [MSSS(τ)=0], and forecasts performing worse than the climatological forecast (i.e. $MSE(\tau)>MSE_C$) have a negative skill score (MSSS(τ)<0).

It is common practice [Lin et al. (2008), Lin and Brunet (2011), Rashid et al. (2011)] to present all MJO forecast verification metrics discussed here as a graph of each metric as function of forecast lead time τ. For positively oriented metrics [e.g., r(τ)], with larger values indicating better forecast performance, such graphs usually display a decreasing curve with large values of the metric for shorter forecast lead times and smaller values for longer forecast lead times. The opposite feature is generally noticed for negatively oriented metrics, with smaller values indicating better forecast performance [e.g. RMSE(τ)]. For these metrics the graphs usually display an increasing curve with small values for shorter forecast lead times and large values for longer forecast lead times.

Finally, it is worth noting that this section addressed MJO forecast verification from a deterministic (ensemble mean) perspective. The reader is encourage to see Marshall et al. (2016) that recently proposed a methodology for probabilistic MJO forecast verification.



(a) (b)

Figure 3: a) Phase space plots of RMM1 and RMM2 computed from NCEP/NCAR reanalysis and satellite OLR (*blue*) and the ensemble-mean POAMA hindcast initialized on 1 January 1986 (*red*), for the period 1 January to 10 February 1986 as shown in Figure 3 of Rashid et al. (2011). The *black dots* are every 5 days. Each octant of the phase diagram is numbered (from 1 to 8) according to the phase definitions of Wheeler and Hendon (2004) Also labelled are the approximate locations of the enhanced convective signal of the MJO for that location of the phase space, e.g., the “Indian Ocean” for phases 2 and 3. The RMM1 and RMM2 values were smoothed with a 1–2–1 filter in time prior to plotting. b) Schematic representation of a MJO forecast F(t,τ) in the RMM1 versus RMM2 phase space for a particular time t produced τ days in advance (i.e. with lead time of τ days) with the corresponding observed MJO signal O(t). See text for additional explanation.

6. Summary, challenges and recommendations in S2S verification

This chapter presented an overview of forecast verification methods relevant to S2S, including current practices. Deterministic and probabilistic verification metrics commonly used for weather and seasonal forecast verification are also used for sub-seasonal forecast verification. However, a number of challenges still need to be addressed including the following:

- Advancing seamless verification practice to allow a smooth comparative quality assessment across different time scales (Zhu et al., 2014; Wheeler et al., 2017);
- Dealing with different ensemble sizes in S2S retrospective forecasts, which are usually much reduced, and real time forecasts when computing forecast probabilities and verification scores (Weigel et al., 2008);
- Advancing the treatment of observational uncertainty in S2S verification (Bellprat et al., 2017);
- Application of spatial verification methods in generally coarse resolution S2S models.

Below are some recommendations for advancing S2S forecast verification research and practice:

- Identify the most relevant forecast quality attributes for the target audience and verification question of interest and choose the appropriate scores for a thorough assessment;
- Develop an S2S forecast verification framework for comparing real time and retrospective forecast skill levels; In the light of the richness of the S2S project database (Vitart et al. 2017) in terms of available retrospective forecasts and near real time forecasts from several modeling centers, and the need for the production of verification information in support of future routine sub-seasonal forecast delivery, there is clearly the need for producing verification information to help forecasters and users from various sectors to acquire knowledge about the strengths and weaknesses of these forecasts, in order to build confidence on S2S forecast products (Coelho et al. 2018);
- Use verification metrics meaningful to users (e.g., use user-relevant thresholds when verifying probabilistic forecasts);
- Move beyond traditional weekly/fortnightly verification toward more user oriented procedures (e.g., active and break rainfall phases, dry/wet spells, heat wave forecast verification). Various application sectors usually require detailed weather within climate information, which are not traditionally verified. The S2S project database (Vitart et al. 2017) provides an excellent opportunity to assess forecast quality of these longstanding demands of various sectors;
- Use appropriate verification measures when dealing with extreme events (e.g., Stephenson et al. 2008; Ferro and Stephenson 2011) such as, heat waves, cold snaps, droughts, and extended rainy conditions;
- Use novel verification measures adequate for S2S forecasts (e.g., probabilistic measures such as the Generalized discrimination score [Weigel et al. 2008, Weigel and Mason 2011] and spatial methods that provide performance information for forecasts with coherent structures [Gilleland et al. 2009] if the spatial resolution of the forecasts allows such detailed spatial verification;
- Explore the novel concept of fair scores in S2S forecast verification (Fricker et al. 2013, Ferro 2014);
- Address sampling uncertainties when computing scores using, for example, bootstrap procedures (Doblas-Reyes et al. 2009) for generating verification measures confidence intervals and producing statistically meaningful comparisons between forecasting systems. Due to the generally limited number of available S2S retrospective forecasts and near real time forecasts it becomes important to have strategies for estimating the uncertainties around the computed verification scores. The bootstrap procedure, which allows the computation of a large number of verification scores by re-sampling the limited number of available forecasts, is an interesting alternative for this purpose;
- Further explore the framework for probabilistic two-dimensional phase space MJO forecast verification (Marshall et al. 2016). Until very recently, MJO forecast verification has been performed using deterministic scores based on ensemble mean forecasts. Again, the S2S project database (Vitart et al. 2017), which contains a very rich amount of ensemble

retrospective forecasts and near real time forecasts from various modeling centers, provides an excellent opportunity for advancing probabilistic MJO forecast verification practice;

- Advance conditional verification practices such as verification conditional on, for example, the MJO, and the El Niño Southern Oscillation (ENSO) phases as well as on particular weather regimes. As MJO and ENSO are recognized as important predictability sources on the S2S time scale, more studies aiming to diagnose the impact of these two phenomena on the prediction ability of current S2S models in variables such as precipitation and near surface temperature, among several others, are required.

References

Baddeley AJ, 1992a: Errors in binary images and an Lp version of the Hausdorff metric. NieuwArch. Wiskunde, 10, 157–183.

Baddeley AJ, 1992b: An error metric for binary images. Robust Computer Vision: Quality of Vision Algorithms, W. Forstner and S. Ruwiedel, Eds., Wichmann, 59–78.

Bellprat O, Massonnet F, Siegert S, Prodhomme C, Macias-Gomez M, Guemas V and Doblas-Reyes FJ, 2017: Exploring observational uncertainty in verification of climate model predictions. Remote Sensing of the Environment. Under review.

Candille G, Côté C, Houtekamer PL and Pellerin G, 2007: Verification of an ensemble prediction system against observations. *Mon. Wea. Rev.*, **135**, 1140-1147.

Casati B, 2010: New developments of the intensity-scale technique within the Spatial Verification Methods Intercomparison Project. *Wea. Forecasting*, **25**, 113-143.

Coelho, C.A.S., Firpo, M.A.F., de Andrade, F.M., 2018. A verification framework for South American sub-seasonal precipitation predictions. Meteorol. Z. https://doi.org/10.1127/metz/2018/0898.

Davis C., Brown B and Bullock R, 2006a: Object-based verification of precipitation forecasts. Part I: Methods and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134**, 1772-1784.

Davis CA, Brown BG and Bullock RG, 2006b: Object-based verification of precipitation forecasts, Part II: Application to convective rain systems. *Mon. Wea. Rev.* **134**, 1785-1795.

Denis B, Côté J and Laprise R, 2002: Spectral decomposition of two-dimensional atmospheric fields on limited-area domains using the discrete cosine transform (DCT). *Mon. Wea. Rev.*, **130**, 1812-1829.

Denis B, Laprise R and Caya D, 2003: Sensitivity of a regional climate model to the resolution of the lateral boundary conditions. Climate Dyn.,20, 107–126.

Doblas-Reyes FJ, Weisheimer A, Déqué M, Keenlyside N, McVean M, Murphy JM, Rogel P, Smithd D and Palmer TN, 2009: Addressing model uncertainty in seasonal and annual dynamical ensemble forecasts. *Q. J. R. Meteorol. Soc.* 135: 1538–1559

Dubuisson M-P and Jain AK, 1994: A Modified Hausdorff Distance for Object Matching. Proc. International Conference on Pattern Recognition, Jerusalem, Israel, pp 566-568.

Dukhovskoy DS, Ubnoske J, Blanchard-Wrigglesworth E, Hiester HR and Proshutinsky A., 2015: Skill metrics for evaluation and comparison of sea ice models, J. Geophys. Res. Oceans, 120, 5910–5931, doi:10.1002/2015JC010989

Ebert EE, 2008: Fuzzy verification of high resolution gridded forecasts: A review and proposed framework. *Meteorol. Appl.*, **15**, 51-64.

Ebert EE and McBride JL, 2000: Verification of precipitation in weather systems: Determination of systematic errors. *J. Hydrology*, **239**, 179-202.

Ferro CAT, 2007: Comparing probabilistic forecasting systems with the Brier score. *Weather and Forecasting,* **22**, 1076-1088.

Ferro CAT, Richardson DS and Weigel AP, 2008: On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorological Applications*, **15**, 19-24, doi:10.1002/met.45.

Ferro CAT and Stephenson DB, 2011: Extremal Dependence Indices: improved verification measures for deterministic forecasts of rare binary events. *Weather and Forecasting*, 26, 699-713, doi:10.1175/WAF-D-10-05030.1.

Ferro CAT, 2014: Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 140, 1917-1923, doi: 10.1002/qj.2270

Fricker TE*,* Ferro CAT and Stephenson DB, 2013: Three recommendations for evaluating climate predictions*.* Meteorol. Appl. 20*:* 246–255*, doi:* 10.1002/met.1409*.*

Germann U and Zawadzki I, 2004: Scale-dependence of the predictability of precipitation from continental radar images. Part II: Probability forecasts. *J. Appl. Meteorol.* **43**: 74–89.

Gilleland E, Ahijevych D, Brown BG, Casati B and Ebert E, 2009: Intercomparison of spatial forecast verification methods. *Weather and Forecasting*, **24**, 1416-1430.

Gilleland E, Lindstrom J and Lindgren F, 2010: Analyzing the image warp forecast verification method on precipitation fields from the ICP. *Wea. Forecasting*, **25**, 1249-1262.

Gilleland E, 2011: Spatial forecast verification: Baddeley’s delta metric applied to the ICP test cases. *Wea. Forecasting*, **26**, 409-415.

Gottschalck J, Wheeler M, Weickmann K, Vitart F, Savage N, Lin H, Hendon H, Waliser D, Sperber K, Nakagawa M, Prestrelo C, Flatau M, Higgins W, 2010: A Framework for Assessing Operational Madden–Julian Oscillation Forecasts: A CLIVAR MJO Working Group Project. *Bull. Amer. Meteor. Soc.*, **91**, 1247–1258.

Hamill TM and Juras J, 2006: Measuring forecast skill: is it real skill or is it the varying climatology? *Q. J. Royal Met. Soc.*, **132**, 2905-2923.

Hudson D, Alves O, Hendon HH, Marshall AG, 2011: Bridging the gap between weather and seasonal forecasting: Intraseasonal forecasting for Australia. Quart. J. Roy. Meteor. Soc., 137,673–689, doi:10.1002/qj.769 http://onlinelibrary.wiley.com/doi/10.1002/qj.769/abstract

Hudson D, Marshall AG, Yin Y, Alves O, Hendon HH, 2013: Improving intraseasonal prediction with a new ensemble generation strategy. Mon Wea Rev*,* 141, 4429-4449. doi: http://dx.doi.org/10.1175/MWR-D-13-00059.1

Jolliffe, I. and Stephenson DB, 2012: *Forecast verification: A practitioner's guide* (I. Jolliffe and D. Stephenson, editors), Wiley, 292 pp.
Jung T and Leutbecher M, 2008: Scale-dependent verification of ensemble forecasts. *Quart. J. Royal Meteorol. Soc.*, **132**, 2905-2923.

Keil C and Craig GC, 2007: A displacement-based error measure applied in a regional ensemble forecasting system. *Mon. Wea. Rev.*, **135**, 3248-3259.

Keil C and Craig GC, 2009: A displacement and amplitude score employing an optical flow technique. *Wea. Forecasting*, **24**, 1297-1308.

Lemieux J-F, Beaudoin C, Dupont F, Roy F, Smith GC, Shlyaeva A, Buehner M, Caya A, Chen J, Carrieres T, Pogson L, DeRepentigny P, Plante A, Pestieau P, Pellerin P, Ritchie H, Garric G and Ferry N, 2005. The Regional Ice Prediction System (RIPS): verification of forecast sea ice concentration. *Q. J. R. Meteorol. Soc.* 142 (695): 632–643

Lin H, Brunet G, Derome J, 2008: Forecast Skill of the Madden–Julian Oscillation in Two Canadian Atmospheric Models. *Mon. Wea. Rev.*, **136**, 4130–4149.

Lin H and Brunet G, 2011: Impact of the North Atlantic Oscillation on the forecast skillof the Madden- Julian Oscillation. GEOPHYSICAL RESEARCH LETTERS, VOL. 38, L02802, doi:10.1029/2010GL046131

Livina V, Edwards N, Goswami S, Lenton T, 2008: A wavelet-coefficient score for comparison of two-dimensional climatic-data fields. Quarterly Journal of the Royal Meteorological Society 134(633): 941–955.

Madden R and Julian P, 1971: Detection of a 40–50-day oscillation in the zonal wind in the tropical Pacific. J. Atmos. Sci., 28, 702–708.

Madden R and Julian P, 1972: Description of global-scale circulation cells in the tropics with a 40–50-day period. J. Atmos. Sci., 29, 1109–1123.

Madden R and Julian P, 1994: Observations of the 40–50-day tropical oscillation: A review. Mon. Wea. Rev., 112 ,814 – 8 37.

Marshall AG, Hendon HH and Hudson D, 2016: Visualizing and verifying probabilistic forecasts of the Madden-Julian Oscillation. Geophysical Research Letters. 43(23), 12278–12286. DOI: 10.1002/2016GL071423

Mason I, 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **30**, 291-303.

Mason SJ and Graham NE, 2002: Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation, *QJRMS* **128**, 2145-2166.

Marsigli C, Boccanera F, Montani A, Paccagnella T., 2005: The COSMO – LEPS ensemble system: validation of the methodology and verification. Nonlinear Processes in Geophysics 12: 527 – 536.

Marzban C and Sandgathe S, 2010: Optical flow for verification. Wea. Forecasting, 25, 1479 - 1494.

Müller WA, Appenzeller C, Doblas-Reyes FJ and Liniger MA, 2005: A debiased ranked probability skill score to evaluate probabilistic ensemble forecasts with small ensemble sizes. *Journal of Climate*, **18**, 1513–1523.

Murphy AH, 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. Wea. Forecasting, 8, 281-293.

Murphy AH, 1996: The Finley Affair: A signal event in the history of forecast verification. Weather and Forecasting, 11, 4-20.

Murphy AH, 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. Mon. Wea. Rev., 116, 2417–2424.

Murphy, AH and Winkler RL, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115**, 1330-1338.

Nehrkorn T, Hoffman RN, Grassotti C and Louis J-F, 2003: Feature calibration and alignment to represent model forecast errors: Empirical regularization. *Q. J. R. Meteorol. Soc.*, **129**, 195-218.

Park YY, Buizza R, Leutbecher M. 2008. TIGGE: preliminary results on comparing and combining ensembles. *Q. J. R. Meteorol. Soc.* **134**: 2029–2050.

Rashid HA, Hendon HH, Wheeler MC, Alves O, 2011: Prediction of the Madden–Julian oscillation with the POAMA dynamical prediction system. Climate Dynamics. Volume 36, Issue 3-4, pp 649-661

Richardson DS, 2001: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quarterly Journal of the Royal Meteorological Society*, **127**, 2473–2489.

Roberts NM and Lean HW, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78-97.

Roebber PJ, 2009: Visualizing Multiple Measures of Forecast Quality. Weather and Forecasting. **24**, 601-608.

Schwedler BRJ and Baldwin ME, 2011: Diagnosing the Sensitivity of Binary Image Measures to Bias, Location, and Event Frequency within a Forecast Verification Framework. Weather and Forecasting. 26: 1032-1044.

Stephenson DB, Casati B, Ferro CAT and Wilson CA, 2008: The extreme dependency score: a non–vanishing measure for forecasts of rare events. *Meteorol Appl* 15:41–50

Taylor KE, 2001: Summarizing multiple aspects of model performance in a single diagram. *JGR*, **106**, 7183–7192

Theis SE, Hense A and Damrath U, 2005: Probabilistic precipitation forecasts from a deterministic model: a pragmatic approach. *Meteorol. Appl.*, **12**, 257-268.

Vigaud N, Robertson AW and Tippett MK, 2017a: Multimodel Ensembling of Subseasonal Precipitation Forecasts over North America. Mon. Wea. Rev., **145**, 3913-3928. DOI: 10.1175/MWR-D-17-0092.1

Vigaud N, Robertson AW, Tippett MK and Acharya N, 2017b: Subseasonal predictability of Boreal Summer Monsoon Rainfall from Ensemble Forecasts., Front. Environ. Sci. 5:67. DOI: 10.3389/fenvs.2017.00067

Vincent LA and Mekis É, 2006: Changes in Daily and ExtremeTemperature and Precipitation Indices for Canada over the Twentieth Century, Atmosphere-Ocean, 44:2, 177-193, DOI: 10.3137/ao.440205

Vitart F, Ardilouze C, Bonet A, Brookshaw A, Chen M, Codorean C, Déqué M, Ferranti L, Fucile E, Fuentes M, Hendon H, Hodgson J, Kang H-S, Kumar A, Lin H, Liu G, Liu X, Malguzzi P, Mallas I, Manoussakis M, Mastrangelo D, MacLachlan C, McLean P, Minami A, Mladek R, Nakazawa T, Najm S, Nie Y, Rixen M, Robertson AW, Ruti P, Sun C, Takaya Y, Tolstykh M, Venuti F, Waliser D, Woolnough S, Wu T, Won D-J, Xiao H, Zaripov R, and Zhang L, 2017: The subseasonal to seasonal (S2S) prediction project data base. BAMS, Jan 2017, 163-173. http://journals.ametsoc.org/doi/pdf/10.1175/BAMS-D-16-0017.1

Vitart F and Molteni F, 2010: Simulation of the Madden- Julian Oscillation and its teleconnections in the ECMWF forecast system. *Quarterly Journal of the Royal Meteorological Society* **136**:10.1002/qj.v136:649, 842-855. http://onlinelibrary.wiley.com/doi/10.1002/qj.623/pdf

Weigel AP, Liniger MA and Appenzeller C, 2007. The discrete Brier and ranked probability skill scores. *Mon. Weather Review*, **135**, 118–124.

Weigel A, Baggenstos D, Liniger MA, Vitart F and Appenzeller C, 2008: Probabilistic verification of monthly temperature forecasts. *Mon. Weather Review* **136**: 5162–5182. doi: http://dx.doi.org/10.1175/2008MWR2551.1

Weigel AP and Mason S, 2011: The Generalized Discrimination Score for ensemble forecasts. *Monthly Weather Review*, **139**, 3069-3074.

Wheeler MC, Zhu H, Sobel AH, Hudson D, Vitart F, 2017: Seamless precipitation prediction skill comparison between two global models. QJRMS., 143 (702), 374–383.

Wheeler MC and Hendon HH, 2004: An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. Mon . Wea . Rev., 132, 1917–1932.

Wilks DS, 2011: Statistical Methods in the Atmospheric Sciences. 3rd Edition. Elsevier, 676 pp.

Wilson L, 2000: Comments on “Probabilistic prediction of precipitation using the ECMWF ensemble prediction system”. *Wea. Forecasting*, **15**, 361-364.

Wilson L, 2014: Forecast Verification for the African Severe Weather Forecasting Demonstration Projects, WMO Technical Document TD 1132, 38pp. *Available on the WMO website.*

Wilson L. and Giles A, 2013: A new index for the verification of accuracy and timeliness of weather warnings. *Meteorol. Appl*., **20**, 206-216

Yates E, Anquetin S, Ducrocq V, Creutin J-D, Ricard D and Chancibault K, 2006: Point and areal validation of forecast precipitation fields. *Meteorol. Appl.*, **13**, 1-20.

Zhang C, 2005: Madden–Julian oscillation. Rev. Geophys., 43, 1–36.

Zhu H, Wheeler MC, Sobel AH, Hudson D (2014) Seamless Precipitation Prediction Skill in the Tropics and Extratropics from a Global Model. Mon. Wea. Rev., 142, 1556–1569.