**Climate Dynamics manuscript No.**
(will be inserted by the editor)

# Improving subseasonal precipitation forecasts through a statistical-dynamical approach : application to the southwest tropical Pacific

**Damien Specq** · **Lauriane Batté**

Received: date / Accepted: date

**Abstract** Subseasonal forecasts are based on coupled general circulation models that often have a good representation of large-scale climate drivers affecting rainfall. Yet, they have more difficulty in providing accurate precipitation forecasts. This study proposes a statistical-dynamical post-processing scheme based on a bayesian framework to improve the quality of subseasonal forecasts of weekly precipitation. The method takes advantage of dynamically-forecast precipitation (calibration) and large-scale climate features (bridging) to enhance forecast skill through a statistical model. It is applied to the austral summer precipitation reforecasts in the southwest tropical Pacific, using the Météo-France and ECMWF reforecasts in the Subseasonal-to-seasonal (S2S) database. The large-scale predictors used for bridging are climate indices related to El Niño Southern Oscillation and the Madden-Julian Oscillation, that are the major sources of predictability in the area. Skill is assessed with a Mean Square Skill Score for deterministic forecasts, while probabilistic forecasts of heavy rainfall spells are evaluated in terms of discrimination (ROC skill score) and reliability. This bayesian method leads to a significant improvement of all metrics used to assess probabilistic forecasts at all lead times (from week 1 to week 4). In the case of the Météo-France S2S system, it also leads to strong error reduction. Further investigation shows that the calibration part of the method, using forecast precipitation as a predictor, is necessary to achieve any improvement. The bridging part, and particularly the ENSO-related information, also provides additional discrimination skill, while the MJO-related information is not really useful beyond week 2 over the region of interest.

**Keywords** Subseasonal prediction · Bayesian statistical post-processing · Calibration · Bridging · El Niño Southern Oscillation · Madden-Julian Oscillation

## 1 Introduction

Subseasonal forecasts (from two weeks to two months) have met growing interest in the last few years, following the launch of the WWRP/WCRP Subseasonal-to-seasonal (S2S)

D. Specq · L. Batté
Centre National de Recherches Météorologiques
Université de Toulouse, Météo-France, CNRS
42, avenue Gaspard Coriolis 31000 Toulouse, FRANCE
E-mail: damien.specq@meteo.fr

prediction project (Vitart et al., 2017). Although it is usually considered a challenging time scale due to the chaotic nature of the atmosphere (Vitart, 2004), recent modeling progress combined with a better understanding of S2S sources of predictability have led to significant improvements in forecast skill (Vitart, 2014).

These sources of predictability include atmospheric boundary conditions varying slowly at subseasonal time scales, such as sea surface temperatures, soil moisture, snow cover and sea ice extent (Robertson and Vitart, 2019). They also include atmospheric waves and oscillations, the most important being the Madden-Julian Oscillation (MJO, Zhang, 2013). Finally, stratospheric processes, such the Quasi-Biennal Oscillation (QBO, Lim et al., 2019) and Sudden Stratospheric Warmings (SSW, Karpechko et al., 2018) are also involved. Under certain conditions, called ”windows of opportunity”, these sources of predictability may enhance prediction skill, not only locally but also remotely through atmospheric teleconnections, e.g active MJO conditions enhancing North Atlantic Oscillation (NAO) predictability (Vitart, 2014).

Consequently, a large number of studies show that S2S systems skillfully predict some meteorological variables at weeks 3 and 4, for instance temperature (e.g Liang and Lin, 2018; Tian et al., 2017; Mastrangelo and Malguzzi, 2019; Wang and Robertson, 2019; Pegion et al., 2019), the MJO (Vitart, 2017; Kim et al., 2018; Marshall and Hendon, 2019) and the NAO (Vitart, 2014). On the other hand, although it may extend beyond weather time scales, the ability to predict precipitation is often lower than that of prognostic variables such as temperature. This has been illustrated for a wide range of geographical areas, e.g North America (Vigaud et al., 2017a), the contiguous United States (Tian et al., 2017), monsoonal regions (Vigaud et al., 2017b), East Africa (Vigaud et al., 2018), East Asia (Liang and Lin, 2018), South America (Coelho et al., 2018), Australia (Hudson et al., 2011; Marshall et al., 2011; Marshall and Hendon, 2015), the southwest tropical Pacific (Specq et al., 2020, in press), and more generally at a global scale (de Andrade et al., 2018; Mastrangelo and Malguzzi, 2019). Yet, indications of significant skill can be found beyond week 2 when considering specific regions, such as the equatorial Pacific where precipitation is directly constrained by the state of El Niño Southern Oscillation (ENSO, de Andrade et al., 2018), or when considering specific rainfall indicators, such as the monsoon onset date (Bombardi et al., 2017) and the monthly accumulated precipitation (Tian et al., 2017).

Since S2S systems are generally able to forecast some large-scale climate features satisfactorily beyond short-range time scales (e.g Vitart, 2017), and since these climate features are known to impact precipitation patterns in wide areas of the globe, it can be assumed there is room for improvement of subseasonal precipitation forecasts through statistical post-processing. Indeed, as pointed out by Schepen et al. (2014) for the seasonal time scale, large-scale climate features in GCMs (General Circulation Models) can be partly disconnected to rainfall, because rainfall is influenced by processes that need to be parameterized as they occur at a much smaller spatial scale than the GCMs resolution. Similarly, S2S rainfall ensemble forecasts generally exhibit biases related to systematic errors and are poorly reliable (in terms of probabilities) beyond week 2 (e.g Vigaud et al., 2017a).

On the other hand, there is a growing demand of subseasonal forecasts among a wide variety of users, such as the energy, agriculture, water management, finance and insurance sectors (Hudson et al., 2011; White et al., 2017). These users obviously require skillful, reliable and bias-free forecasts. As a result, statistical correction and post-processing of S2S forecasts is a crucial matter, in particular for a parameterized variable like rainfall. Two general and complementary approaches can be developed for this post-processing: calibration and bridging (Schepen et al., 2014). A statistical-dynamical post-processing scheme is called calibration when the predictor for the observed variable (i.e precipitation) is the raw

variable itself as forecast by the dynamical forecasting system. It is referred to as bridging when the predictors are large-scale climate features, that are also forecast by the dynamical system. The link between predictors and the predicted variable is established through a statistical model that may have the same mathematical formulation for both calibration and bridging.

Calibration and bridging have been jointly applied in a large number of studies related to seasonal forecasting of temperature or precipitation (e.g Schepen et al., 2012, 2014, 2016; Strazzo et al., 2019), and recent studies have started carrying out this work for S2S time scales. In terms of calibration-oriented approaches for precipitation, Li et al. (2019) have proposed a bias correction methodology for S2S precipitation at the scale of hydrological catchments, while Vigaud et al. (2019) have proposed a spatial correction methodology for multimodel subseasonal precipitation forecasts using local Laplacian eigenfunctions. Likewise, Doss-Gollin et al. (2018) have tested several Model Output Statistics (MOS) methods for S2S precipitation prediction over central South America that also lead to better calibration. On the other hand, similar to the bridging approach, statistical-only subseasonal forecasts based on large-scale climate features have been developed using machine learning (Cohen et al., 2019, for temperature and precipitation) or empirical models (Johnson et al. 2014 for temperature, Baggett et al. 2018 for tornadoes and hail).

In this study, we introduce a simple bayesian statistical-dynamical scheme dealing with both calibration and bridging in the same framework. This scheme is applied to precipitation forecast in the southwest tropical Pacific (SWTP) with the Météo-France and ECMWF S2S systems. It uses ENSO and MJO-related predictors for bridging, along with the forecast precipitation at grid point level for calibration. The SWTP domain ranges from 110°E to 200°E in longitude, and 30°S to 0° in latitude. Rainfall in this area is a crucial matter with the regular occurrence of tropical cyclones and heavy rainfall spells (McGree et al., 2014) but also droughts (McGree et al., 2016), affecting Pacific island territories as well as Australia. The focus is put on the austral summer season (December, January and February) as it corresponds to the peak season for MJO activity in the Southern Hemisphere (Zhang, 2005), to the wet season in southwest Pacific island territories and to the monsoon season in Northern Australia (Marshall and Hendon, 2015). It has been widely demonstrated that precipitation patterns in the area are strongly influenced both by ENSO, as a slowly-varying process for the subseasonal time scales, and by the MJO (e.g de Andrade et al., 2018). Since these phenomena are recognized to be well represented in subseasonal forecasting systems, using them as predictors in the statistical-dynamical scheme should provide additional information and improve precipitation forecast skill compared to the raw precipitation output from the numerical models.

This article is structured as follows. Section 2 is dedicated to the description of the S2S reforecasts and the rest of the data used for this study. Section 3 goes into details of the statistical-dynamical approach implemented here and introduces the framework for forecast verification and comparison. Section 4 presents the main improvements obtained with the statistical-dynamical scheme in terms of forecast skill and reliability compared to the raw S2S forecasts, while Section 5 discusses the relative importance of each large-scale predictor used for bridging. Finally, Section 6 summarizes and discusses the main results of the article.

**Table 1** Reforecast attributes for the Météo-France and ECMWF systems

| Attributes | ECMWF | MF |
|---|---|---|
| Time range | Day 1-46 | Day 1-61 |
| Atmospheric resolution | T639/319 L91 | T255L91 |
| Reforecast | On the fly | Fixed |
| Reforecast period | Past 20 yrs (1996-2015) | 1993-2014 |
| Reforecast frequency | Two per week | Four per month |
| Ensemble number | 11 | 15 |
| Coupling | Ocean and sea ice | Ocean and sea ice |

**Table 2** Correspondence between the Météo-France and the ECMWF reference start dates

| **MF** | 12-01 | 12-08 | 12-15 | 12-22 | 01-01 | 01-08 | 01-15 | 01-22 | 02-01 | 02-08 | 02-15 | 02-22 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ECMWF** | 12-01 | 12-08 | 12-15 | 12-22 | 12-31 | 01-07 | 01-14 | 01-21 | 02-01 | 02-08 | 02-15 | 02-22 |

## 2 Data

### 2.1 S2S reforecasts

The ensemble reforecasts from Météo-France (MF) and the European Centre for Medium-range Weather Forecasts (ECMWF) were extracted from the S2S database (Vitart et al., 2017). These two systems are both based on coupled GCMs and their main features are summarized in Table 1. Their output data was extracted on a common 1.5° grid on the SWTP domain (110°E-200°E; 30°S-0°). The 1.5° resolution corresponds to the common initial archiving resolution of the S2S database and has already been adopted in other S2S studies (e.g Vigaud et al., 2017a).

In order to make a proper comparison between these two systems, it is essential to choose a common reforecast period for verification and implementation of the statistical-dynamical approach. The MF system uses a fixed reforecast period (1993-2014) while the ECMWF system updates its reforecast for every real-time forecast by running the system at the same calendar date over the past 20 years. We therefore chose to consider the 19-year reforecast period 1996-2014, using the 2016 version of the ECMWF reforecasts.

In addition, we restrained our selection to the start dates in the DJF season. However, these start dates also differ between the two systems (e.g MF starts on January 8 while ECMWF starts on January 7) and ECMWF has actually more start dates (two per week) than MF does (four per month). As we chose to keep all DJF start dates from the MF reforecast, we only retained the closest corresponding start dates in the ECMWF reforecast in order to compare the two systems' performance with the same sample size. This correspondence is specified in Table 2. This leads to 216 start dates (18 years x 3 months x 4 startdates per month) under consideration for each S2S system. Both systems are evaluated separately on weekly windows after discarding the first four days since they belong to short-range forecasting. If $d$ is the day of the start date, the weekly verification calendar includes week 1 ($[d+5, d+11]$) to week 4 ($[d+26, d+32]$).

## 2.2 Precipitation data

S2S precipitation data for MF and ECMWF is extracted on the 1.5° grid over the domain 110°E-200°E; 30°S-0° (60 x 21 grid points) and averaged on weekly windows according to the procedure described in Section 2.1. It is verified against a reference precipitation dataset called Multi-Source Weighted-Ensemble Precipitation (MSWEP) version 1.2 (Beck et al., 2017). This dataset is available for the 1979-2015 period and covers the whole globe on a 0.25° grid. It combines various precipitation sources: satellite data, World Meteorological Organization Global Telecommunication System (WMO GTS) rain gauges and reanalysis. MSWEP data was interpolated on the 1.5° reforecast grid through conservative remapping, and temporally averaged along the same weekly windows as the reforecasts.

## 2.3 Large-scale predictors

In this study, we considered the two most widely recognized sources of subseasonal predictability for precipitation in the SWTP, ENSO and the MJO. We decided to characterize the ENSO state at a given day by the sea surface temperature anomalies, spatially-averaged over the standard Niño 3.4 box (170°W-120°W; 5°S-5°N), and temporally-averaged over the previous 90 days. This index will be noted $N34$ hereafter. Other ENSO-related indices were tested but showed little difference for the purpose of our statistical-dynamical scheme. As for the MJO, it is represented by the two indices $RMM1$ and $RMM2$ defined by Wheeler and Hendon (2004) and corresponding to the first two principal components of the multivariate principal component analysis carried out on the meridionally-averaged fields of Outgoing Longwave Radiation (OLR) and zonal winds at 850 (U850) and 200 hPa (U200) in the 15°S-15°N equatorial band.

The determination of the indices in the S2S reforecasts first requires to compute anomalies of a given field (SST for $N34$, OLR, U850 and U200 for $RMM1$ and $RMM2$). The anomalies are calculated for every daily lead time by leave-one-year-out cross-validation: for a given start date in a given year and at a given lead, the anomaly is the variable in the reforecast minus the averaged variable at this start date and lead in the reforecasts of all other years. Moreover, the large-scale predictors are also computed in the ERA-Interim reanalysis (Dee et al., 2011) to train the statistical model with real-world data. In order to be consistent between reforecast and reference data, the anomalies in ERA-Interim are computed with the same method by subtracting the average variable for the same calendar day in all other years of the 1996-2014 period. Finally, the computation of the indices in the S2S reforecasts often requires to subtract previous values averaged over a period that is longer than the length of the runs (90 days for $N34$, 120 days for MJO indices). Then, for the $n$-th day of the run, we substract the average value composed of the $n-1$ previous days of the run and the $90-(n-1)$ (respectively $120-(n-1)$) days prior to the start dates in ERA-Interim.

Using this procedure, the MJO indices in the S2S reforecasts are computed following the recommendations of Gottschalck et al. (2010) and the ECMWF technical note available on the S2S website (ECMWF, 2017). As for the daily $N34$ index at day $d$, it is obtained by averaging the spatially-averaged SST anomalies over the 90 days from day $d-89$ to day $d$. Finally, although these indices are initially computed as daily values, they are averaged over the weekly windows of the S2S reforecast calendar (described in Section 2.1) for the implementation of the statistical-dynamical scheme.

# 3 Methods

## 3.1 Statistical-dynamical prediction

### *3.1.1 General framework*

The bayesian approach proposed in this study is similar to that proposed by Coelho et al. (2004) that was also applied and completed by Luo et al. (2007), denoted L07 hereafter. Since this approach was initially designed for seasonal forecasting, we have adapted it to subseasonal forecasting. Let's consider that we intend to forecast the distribution of a predictand variable $y$ conditional upon a set of $n$ predictors $\mathbf{x} = (x_1, x_2, ..., x_n)$, noted $p(y|\mathbf{x})$. According to Bayes' theorem,

$$p(y|\mathbf{x}) = \frac{p(\mathbf{x}|y)p(y)}{p(\mathbf{x})} \ . \tag{1}$$

$p(y)$ is the prior distribution of the predictand $y$ before any information is known from a specific forecast. $p(\mathbf{x}|y)$ is called the likelihood: it expresses how $\mathbf{x}$ is distributed when the predictand $y$ is known. $p(y|\mathbf{x})$ is the updated distribution of $y$ when information about $\mathbf{x}$ is available.

L07 have shown that an explicit formulation of $p(y|\mathbf{x})$ is possible in a gaussian framework under the following assumptions:

1. The prior distribution $p(y)$ is normal: $p(y) \sim \mathcal{N}(\mu_p, \sigma_p^2)$
2. For each predictor $x_i$, $p(x_i|y)$ is a normal distribution determined through a linear regression:

$$p(x_i|y) \sim \mathcal{N}(a_i + b_i y, \sigma_i^2) \ , \tag{2}$$

   where $a_i$ is the intercept, $b_i$ is the slope and $\sigma_i^2$ is the variance of the residuals.
3. The residuals in the linear regressions from Equation (2) are independent from each other.

Then, according to L07, $p(y|\mathbf{x})$ also follows a normal distribution, noted $p(y|x) \sim \mathcal{N}(\mu_t, \sigma_t^2)$, and the parameters of this conditional distribution are expressed as follows:

$$\frac{1}{\sigma_t^2} = \frac{1}{\sigma_p^2} + \Sigma_{i=1}^{n} \frac{b_i^2}{\sigma_i^2} \tag{3}$$

$$\frac{\mu_t}{\sigma_t^2} = \frac{\mu_p}{\sigma_p^2} + \Sigma_{i=1}^{n} \frac{b_i^2}{\sigma_i^2} (\frac{x_i - a_i}{b_i}) \tag{4}$$

### *3.1.2 Scheme implementation*

In this study, the predictors are the weekly-averaged forecast large-scale indices $N34$, $RMM1$, $RMM2$, and the weekly-averaged forecast rainfall $r_f$. The predictand variable is the weekly-averaged observed rainfall $r_o$. Assumption n°3 in the previous Section 3.1.1 suggests that any dependency between the predictors should be removed. For this purpose, since it has been widely shown that cross-timescale interactions between ENSO and the MJO might exist (e.g Doss-Gollin et al., 2018), it should be relevant to use the principal components of

($N34$,$RMM1$,$RMM2$) instead of the original indices. However, on a practical point of view, this approach proves to be less relevant in terms of skill improvement (see Supplementary material Figure S1, as compared to Figure 2). Then, it will not be further applied and the original large-scale indices are kept as predictors.

Furthermore, as in L07, we need to overcome the fact that the rain rates $r_o$ and $r_f$ are non-gaussian. This issue is tackled in a similar manner: the weekly rain rates $r_o$ and $r_f$ are converted into normally distributed variables of mean 0 and variance 1, noted $\hat{r_o}$ and $\hat{r_f}$ hereafter. This transformation is made up of two steps using the equal-quantile method: application of the climatological Cumulative Distribution Function (CDF) of rain rates, followed by the inverse CDF of the normal distribution $\mathcal{N}(0,1)$.

Such a transformation is always mathematically feasable and leads to a normally-distributed variable most of the time. One limitaton appears in dry areas where the CDF for a 0 mm rain rate is far greater than the 0% quantile, but it only applies to a small fraction of the SWTP domain (continental Australia, see Section 4.2). In such cases, the transformed variable cannot be fully gaussian as it does not cover the full range of the normal distribution. For instance, if 0 mm represents the lower 20% of the climatology, the transformation will necessarily lead to values greater than the 20th percentile of the $\mathcal{N}(0,1)$ distribution (i.e $\sim -0.84$).

The climatological CDF of rain rates is determined separately for reforecast and observations with a leave-one-year-out cross-validation approach similar to Vigaud et al. (2017a). In the observations, for a given target week, the same calendar weeks in all other 18 years are considered, along with the week before and the week after. This leads to a climatological sample of $18 \times 3 = 54$ values for the target week, from which an empirical CDF is built. Likewise, the S2S reforecast climatology for a given year, start date and lead time is constructed by taking the reforecasts for the same start date and lead time considering the 18 remaining years and all ensemble members. This leads to a climatological sample of $18 \times n$ values to establish the empirical CDF, where $n$ the ensemble size (15 members for MF and 11 for ECMWF).

The bayesian scheme described in Section 3.1.1 is implemented with $y = \hat{r_o}$ and $\mathbf{x} = (\hat{r_f}, N34, RMM1, RMM2)$. For the sake of robustness, we take the ensemble mean of the forecast predictors, including $\hat{r_f}$, and do not consider the spread of the predictor around the ensemble mean. According to the normalizing transformation from $r_o$ to $\hat{r_o}$, the prior distribution of $\hat{r_o}$ is assumed to be $\mathcal{N}(\mu_p = 0, \sigma_p^2 = 1)$. Equations (3) and (4) are applied to deduce the conditional distribution $p(\hat{r_o}|\hat{r_f}, N34, RMM1, RMM2)$. If needed, the resulting mean $\mu_t$ can be back-transformed to obtain a precipitation forecast in mm, by using the CDF of $\mathcal{N}(0,1)$ followed by the inverse climatological CDF of observations.

### *3.1.3 Uncertainty in the predictors' forecasts*

A major difference between the approach in L07 and ours is the inclusion of large-scale climate features as predictors (bridging), while they only used the forecast precipitation as a predictor for the observed precipitation. As a result, we have included a two-term decomposition of the uncertainty $\sigma_i^2$ in the relationship (defined in Equation (2)) between each predictor $x_i$ and the predictand $y = \hat{r_o}$. There are actually two possible values for predictor $x_i$: its real-world value in the reference data, noted $x_{io}$, and its forecast value, noted $x_{if}$. The benefits of the predictor might be limited by its intrinsic predictive ability, but also by the fact that S2S systems do not forecast it with sufficient quality.

The intrinsic predictive ability is established in the reference data with the linear regression:

$$p(x_{io}|y) \sim \mathcal{N}(a_{io} + b_{io}y, \sigma_{io}^2) \ . \tag{5}$$

The ability of the S2S systems to forecast the predictor $x_i$ is assessed in a second linear regression:

$$p(x_{if}|x_{io}) \sim \mathcal{N}(a_{if} + b_{if}x_{io}, \sigma_{if}^2) \ . \tag{6}$$

The composition of the normal distribution from Equation (5) followed by the normal distribution from Equation (6) leads to the resulting expressions of the parameters $a_i$, $b_i$ and $\sigma_i^2$ from Equation (2):

$$a_i = a_{if} + b_{if}a_{io} \tag{7}$$
$$b_i = b_{if}b_{io} \tag{8}$$
$$\sigma_i^2 = b_{if}^2\sigma_{io}^2 + \sigma_{if}^2 \tag{9}$$

$\sigma_i^2$ is the sum of the uncertainty in the prediction of $x_i$ by the S2S systems (i.e the error), noted $\sigma_{if}^2$, and an uncertainty term which is the intrinsic uncertainty of the predictor $\sigma_{io}^2$, multiplied by $b_{if}$. $b_{if}$ is proportional to the correlation score of the forecast predictor $x_{if}$ with the observed predictor $x_{io}$. When the predictor $x_i$ is rainfall itself, $\sigma_{io}^2 = 0$ and $b_{io} = 1$ so the only source of uncertainty is the forecast error $\sigma_{if}^2$. Both linear regressions from Equations (5) and (6) are fitted for each lead time using leave-one-year-out cross-validation.

### 3.2 Assessing predictor importance

Section 5 discusses the respective role of the each predictor in the predictive ability of the statistical-dynamical model. The importance of each predictor is assessed using likelihood-ratio tests (Buse, 1982) in a forward selection mode. Our predictive model is described by Equation (4). Because the prior probability distribution of $\hat{r_o}$ has mean $\mu_p = 0$, the predictive model can be re-written:

$$\hat{r_o} = \Sigma_{i=1}^4 \sigma_t^2 \frac{b_i^2}{\sigma_i^2}(\frac{x_i - a_i}{b_i}) \ , \tag{10}$$

This model consists in four predictive terms, the added value of which needs to be estimated. Similar to a stepwise forward predictor selection for a classical linear regression (Wilks, 2006), we start from a model where $\hat{r_o}$ is a constant and compare this model with all the other models for which we consider each predictive term $\sigma_t^2 \frac{b_i^2}{\sigma_i^2}(\frac{x_i - a_i}{b_i})$ separately. The most important of the predictive terms is the one which returns the smallest p-value in the likelihood-ratio test when comparing with the constant model. In the next step of the procedure, we start from the new model including the most important predictive term, and we determine which of the remaining terms is the second most predictive, using the same p-value criterion. The procedure ends when all predictive terms have been included and it gives the order of importance of each predictive term. Using a p-value stop criterion, set at 0.05 in this study, also enables to eliminate predictive terms that do not bring additional information.

This procedure is carried out at each grid point with two sets of predictors:

1. Large-scale predictors $N34$, $RMM1$ and $RMM2$ only, in the reanalysis data, considering all lead times together
2. All predictors (including $\hat{r_f}$), as forecast by S2S systems, at each lead time separately

The first option aims at identifying where each predictor influences precipitation in real-world data. The second enables to identify, at each lead time, which predictor actually brings information to the statistical-dynamical approach.

### 3.3 Verification

We use a verification framework with a deterministic and a probabilistic score, computed at every lead time (week 1 to week 4). This framework partly re-uses and adapts the verification procedure of subseasonal weekly accumulated precipitation predictions proposed by (Coelho et al., 2018) and called "all season hindcast verification". Deterministic verification is carried out with the Mean Square Skill Score (MSSS), which is defined as

$$MSSS = 1 - \frac{MSE}{MSE_{clim}} \ , \tag{11}$$

where $MSE$ is the mean square error of the reforecasts, and $MSE_{clim}$ is the mean square error of a climatological forecast that is constructed from the observations by cross-validation similarly to the procedure detailed in Section 3.1.2. When computing the MSSS, a simple cross-validated bias correction is applied to the raw forecasts so that the Mean Square Error does not include any systematic error that would lower the scores without being related to the predictive ability of the systems.

We also consider probabilistic forecasts of a binary event corresponding to weekly precipitation above the climatological upper quintile, because our main objective is to assess and improve the ability of S2S systems to detect spells of heavy rainfall. The forecast probability of the upper quintile in the raw S2S forecasts is evaluated by taking the fraction of ensemble members for which the forecast value exceeds the upper quintile threshold. For the output of the statistical-dynamical scheme, the forecast probability corresponds to the probability that a variable following a normal distribution with the parameters in Equations (3) and (4) exceeds the upper quintile of the $\mathcal{N}(0,1)$ normal distribution (i.e $\sim 0.84$).

The discrimination of the probabilistic forecasts (i.e whether the forecasts correctly make the difference between occurrence and non-occurrence of the binary event) is assessed with the area under the Relative Operating Characteristic (ROC) curve $A$. $A$ is rescaled as the ROC skill score (ROCSS), i.e

$$ROCSS = 2A - 1 \ , \tag{12}$$

so that values are comprised between 0 and 1 for forecasts that are more skillful than climatology. Finally, we also assess the reliability of the probabilistic forecasts using reliability diagrams and the reliability component of the Brier Score (Murphy, 1973).

Moreover, the verification scores MSSS and ROCSS will either be computed as a unique score at the scale of the whole SWTP domain, for the sake of brevity, or at grid point level. When the ROCSS is shown on maps, the value at a given grid point is actually computed by pooling the forecast/observation pairs for a neighborhood of nine grid points ($3 \times 3$) around the central grid point. This enables to reduce spatial noise and to increase robustness of a comparison between the ROCSS of two forecasts.

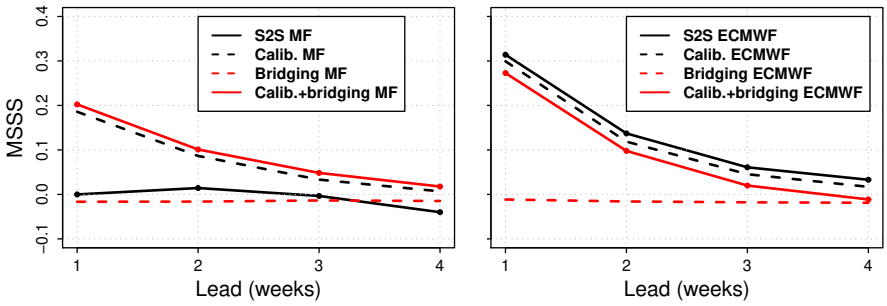


**Fig. 1** Mean Square Skill Score at the scale of the SWTP basin for the subseasonal reforecasts with various statistical-dynamical post-processing: raw S2S (solid black line), calibrated (dashed black line), bridged (dashed red line) and calibrated + bridged (solid red line). Left: MF S2S system. Right: ECMWF S2S system. Note that a cross-validated bias correction has been performed to the raw S2S systems before computing the MSSS.

Such a comparison between forecast $A$ and forecast $B$ is made by taking the difference $ROCSS_B - ROCSS_A$. For this difference to be considered significant, we set two simultaneous conditions:

1. The greater of $ROCSS_A$ and $ROCSS_B$ must be significant at the 95% level according to a Mann-Whitney U test (Wilks 2006)
2. The greater of $ROCSS_A$ and $ROCSS_B$ must be significantly greater than the other one, at the 95% level, according to a one-sided DeLong test (DeLong et al., 1988)

## 4 Results

### 4.1 Comparison over the whole domain

Four types of subseasonal reforecasts of precipitation are assessed in this section: the raw S2S reforecasts, the calibrated reforecasts (forecast precipitation $\hat{r_f}$ is the only predictor), the bridging reforecasts (large-scale predictors $N34$, $RMM1$ and $RMM2$ without forecast precipitation) and the calibration + bridging reforecasts (all predictors $\hat{r_f}$, $N34$, $RMM1$ and $RMM2$). The aim is to have an overview of the impact of each statistical-dynamical approach on reforecast skill. For the sake of brevity, results are shown with skill scores computed by pooling all the grid points in the SWTP domain (see Section 3.3).

Figure 1 illustrates such a comparison for the Mean Square Skill Score, once the mean bias has been removed from the raw S2S reforecasts (solid black line). The first notable result from Figure 1 is that the bridging approach alone leads to the same amount of errors as a climatological forecast ($MSSS = 0$), for both the MF and ECMWF systems. A second notable result is that the raw MF system has apparently such large residual errors that it does not really perform better than climatology, even after the simple bias correction. On the contrary, this is not the case for the raw ECMWF system. A third important point is that all approaches involving calibration exhibit good performances and less errors than climatology up to week 4. This leads to a drastic improvement for the MF system, but not for the ECMWF system. Indeed, the ECMWF system makes actually less error with the simple bias correction than with any statistical-dynamical approach, a result which is presumably due to the already good performance of the raw ECMWF reforecasts in terms

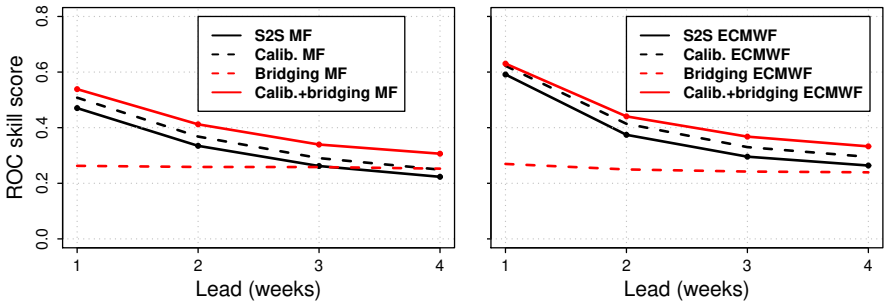


**Fig. 2** ROC skill score for the upper quintile of precipitation at the scale of the SWTP basin for the subseasonal reforecasts with various statistical-dynamical post-processing: raw S2S (solid black line), calibrated (dashed black line), bridged (dashed red line) and calibrated + bridged (solid red line). Left: MF S2S system. Right: ECMWF S2S system. The differences between the solid black line, the dashed black line and the solid red line are always significant at the 95% according to a one-sided DeLong test.

of MSSS. Finally, for both systems, there does not seem to be any added value of calibration + bridging compared to calibration alone, suggesting that bridging does not bring any error correction.

Figure 2 shows the same comparison for the ROC skill score of the upper quintile of precipitation. For this score, results are similar for both S2S systems: the performance of the calibrated reforecasts is better than that of the raw reforecasts, but the calibrated reforecasts are also improved by the calibration + bridging approach, at least for the later lead times (weeks 3 and 4). However, bridging alone leads to an almost constant skill score (no variation with lead) that is less than the scores obtained by any other approach. We find that this constant score actually illustrates a baseline discrimination ability related to the $N34$ predictor. Indeed, this predictor remains almost constant, and very close to the initial reference value, in the first four weeks of the S2S runs (not shown).

The results from Figures 1 and 2 suggest that calibration is useful to error correction to some extent, when the raw forecasts exhibit large errors, and it is also useful to improve the detection of heavy rainfall events. In the meantime, bridging brings additional value for discrimination, provided it is combined with calibration. These results are also confimed on four other S2S systems (see Supplementary material, Figures S2 and S3). Yet, it must be acknowledged that the proposed approach is not able to make up for the skill difference between the input dynamical models, i.e ECMWF remains better than Météo-France even after statistical post-processing.

## 4.2 Comparison of discrimination ability at grid point level

Figures 3 and 4 represent the difference in terms of ROC skill score between the calibration + bridging reforecasts and the raw S2S reforecasts, using the MF and ECMWF systems respectively. The differences are shown alongside the initial skill of the raw S2S system, and the ROC skill scores are computed by pooling the grid points over a $3 \times 3$ grid point neighborhood as explained in Section 3.3. In agreement with Figure 2, the calibration + bridging approach leads to a significant improvement of forecast skill in a large fraction of the domain.

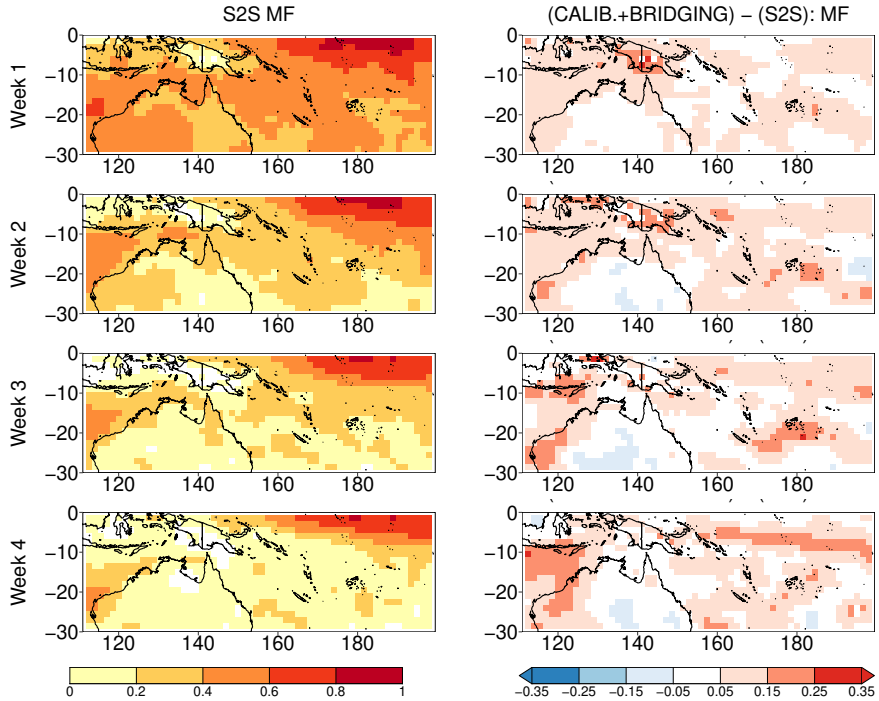


**Fig. 3** Left: ROC skill score for the upper quintile of weekly precipitation in the MF S2S reforecast. Right: Difference in ROC skill score between the calibration + bridging statistical-dynamical scheme and the raw MF S2S reforecast. The ROC skill score are computed by pooling values over a $3 \times 3$ neighborhood. Grid points in white correspond to grid points where the difference is not significant according to the two criteria detailed in Section 3.3.

The benefits of using the statistical-dynamical approach appear to increase with lead time. This is not suprising because the raw reforecasts have more detection failures and are more noisy at later lead times. Consequently, there is more room for improvement through calibration, that corrects distribution errors, and bridging, that dampens the importance of spurious high-frequency variability relative to more relevant lower frequency variability. The main regions of improvement are the Maritime Continent, the southern oceanic part of the domain (including New Caledonia at 166°E-21°S, Vanuatu at 168°E-17°S and Fiji at 178°E-18°S) and the western Australian coast.

One notable exception is a part of continental Australia for which there is no improvement or even a decrease in skill, depending on the lead time. Further investigation (not shown here) found this is related to the dryness of the area, for which the weekly precipitation distribution is highly positively skewed. When transforming precipitation into a gaussian variable, small differences between precipitation values in mm may lead to larger differences in the transformed precipitation. These differences in precipitation are not necessarily related to important differences in large-scale conditions. They lead to non-significant or inconsistent relationships between transformed precipitation and large-scale predictors, such as $N34$, when fitting the linear regression in Equation (5), hence the degraded forecast skill. This is a limitation of our methodology that only affects a restricted part of the chosen domain.

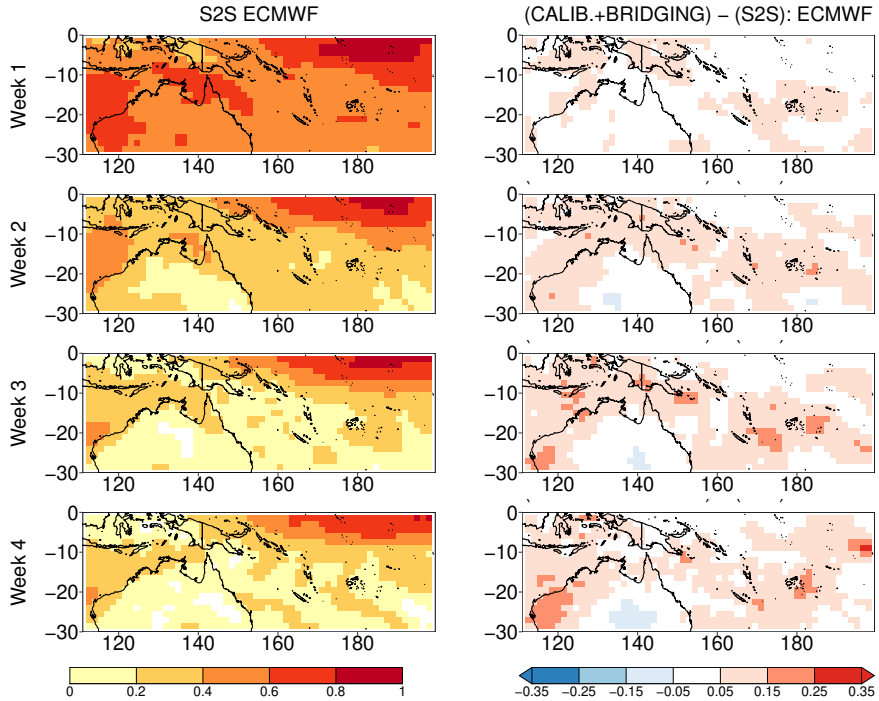


**Fig. 4** Left: ROC skill score for the upper quintile of precipitation in the ECMWF S2S reforecast. Right: Difference in ROC skill score between the calibration + bridging statistical-dynamical scheme and the raw ECMWF S2S reforecast.

## 4.3 Reliability of probabilistic reforecasts

The overall reliability of the raw S2S reforecasts, the calibration and the calibration + bridging reforecasts is evaluated in Figures 5 and 6, at the scale of the whole domain. These diagrams are completed with the reliability and the resolution components of the Brier Score in Table 3. The reliability component is negatively oriented (the smaller the more reliable) while the resolution component is positively oriented. The two statistical-dynamical approaches clearly outperform the raw S2S reforecasts for both reliability and resolution. The result concerning resolution was foreseeable from Figures 2, 3 and 4, given that resolution and discrimination (as assessed by the ROC skill score) are ”two sides of the same coin” (Bröcker, 2015), but the reliability improvement remained to be demonstrated. Moreover, these improvements are notable for all lead times. The comparison between the calibration and calibration + bridging diagrams would at first sight suggest that calibration alone, without bridging, should be preferred because it leads to more reliable forecasts. However, this would happen at the expense of resolution which is better when bridging is added, as could have been inferred from Figure 2. Thanks to the information from large-scale predictors, the calibration + bridging scheme is more skillful at detecting heavy precipitating events but it also tends to be more overconfident.

**Table 3** Reliability and resolution components ($\times 10^2$) of the Brier Score (whole domain) for the upper quintile of weekly precipitation at weeks 1 to 4. Reliability is negatively oriented and resolution is positively oriented.

| | | Week 1 | | Week 2 | | Week 3 | | Week 4 | |
|---|---|---|---|---|---|---|---|---|---|
| | | Rel. | Res. | Rel. | Res. | Rel. | Res. | Rel. | Res. |
| MF | Raw | 1.41 | 2.13 | 0.95 | 1.07 | 0.79 | 0.74 | 0.81 | 0.60 |
| | Calib. | 0.020 | 2.34 | 0.0089 | 1.27 | 0.020 | 0.87 | 0.037 | 0.70 |
| | Calib.+bridging | 0.076 | 2.63 | 0.098 | 1.58 | 0.14 | 1.12 | 0.16 | 0.92 |
| ECMWF | Raw | 0.61 | 3.44 | 0.81 | 1.39 | 0.89 | 0.93 | 0.87 | 0.77 |
| | Calib. | 0.014 | 3.66 | 0.005 | 1.62 | 0.010 | 1.07 | 0.015 | 0.90 |
| | Calib.+bridging | 0.046 | 3.70 | 0.092 | 1.77 | 0.11 | 1.25 | 0.15 | 1.02 |

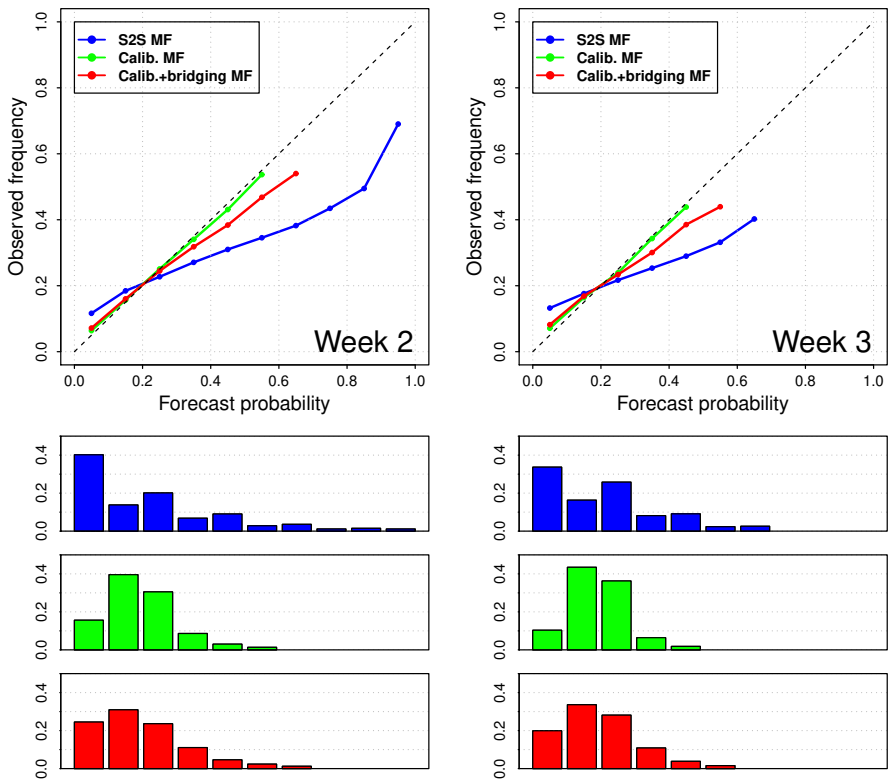


**Fig. 5** Comparison of the reliability diagrams (whole domain) for the upper quintile of weekly precipitation, between the raw MF S2S reforecast (blue), the calibration reforecast (green) and the calibration + bridging reforecast (red), at week 2 (left) and week 3 (right). The forecast frequency in each probability bin is represented below on the bar plots for each of reforecasts, with the same color code. Bins with less than 1% of the total number of forecasts are not plotted.

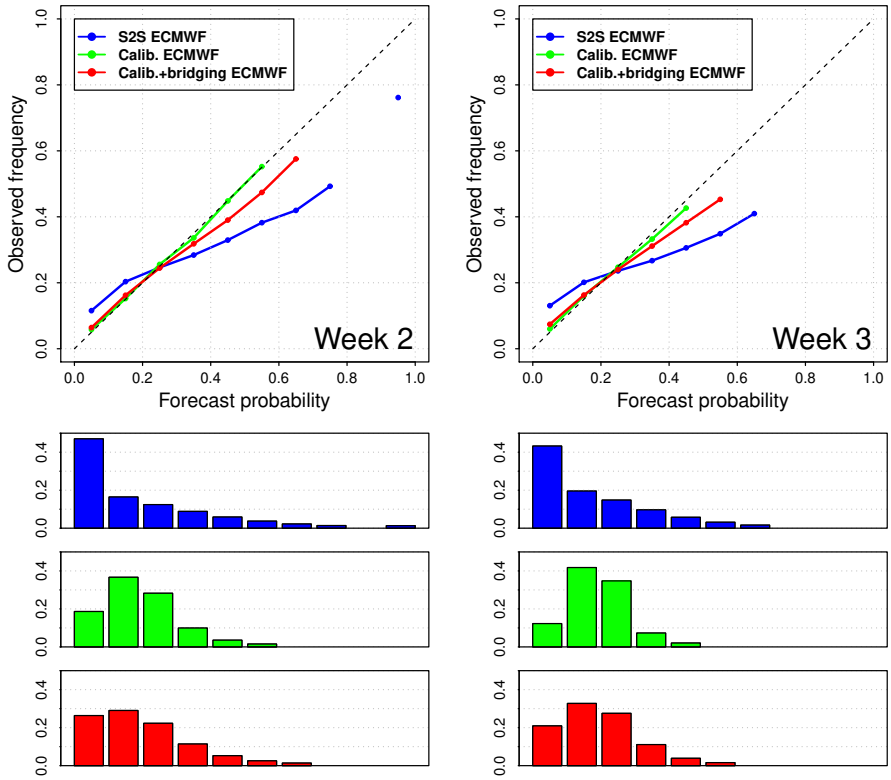


**Fig. 6** Same as Figure 5 for the ECMWF S2S system. Bins with less than 1% of the total number of forecasts are not plotted.

## 5 Role of the large-scale predictors

### 5.1 Theoretical relative added value

Following the methodology described in Section 3.2, the theoretical importance of each predictor $N34$, $RMM1$ and $RMM2$ in the reference datasets is assessed with a stepwise forward selection scheme. For each of these predictors, the first column in Figure 7 shows the coefficient of the linear regression in Equation (5) at each grid point, if it is significant at the 95% level. This is an estimation of the absolute role that each predictor plays individually. The second column in Figure 7 shows the order in which the predictive term in Equation (10), associated to a given predictor, is selected. This is an estimation of the relative role of each predictor in the bridging part of the approach. Grid points in white correspond to grid points where the predictive term is considered not to bring any additional information according to the p-value threshold of 0.05 mentioned in Section 3.2.

The key result is that the $N34$ predictor related to ENSO is the first large-scale source of information over most of the SWTP, mainly in the southern oceanic part and in the north-eastern equatorial region. If not ENSO, the most important predictor is quite often the first MJO index $RMM1$, that is the main source of information in the northwestern part of the domain but is also a significant source of information in a region that extends southeastward

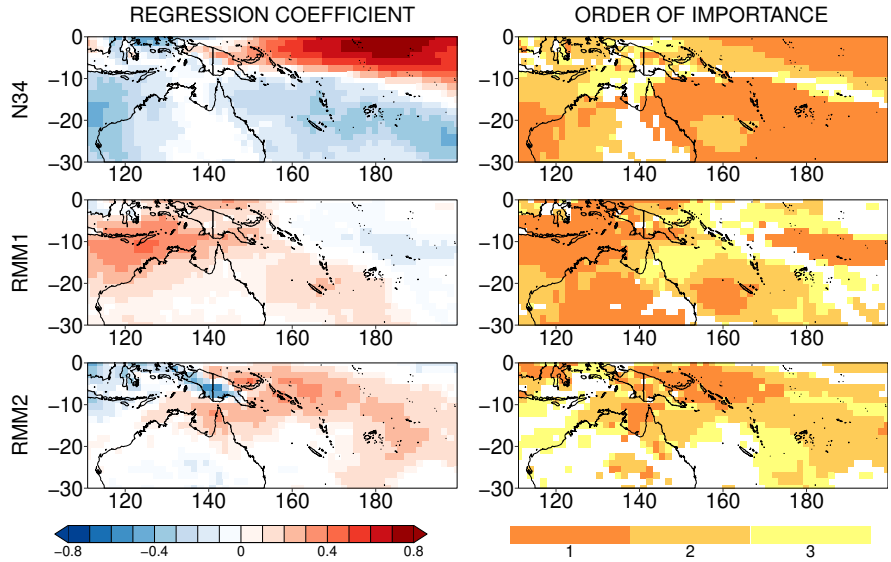


**Fig. 7** Left: linear regression coefficient of each predictor against observed rainfall $\hat{r_o}$. Grid points are in white if the coefficient is not significant at the 95% level.
Right: Order of selection for each predictor in a stepwise forward selection scheme when considering large-scale predictors only in reference data. Grid points in white correspond to grid points where adding the predictor leads to a p-value greater than 0.05 in the likelihood-ratio test.

from the Maritime Continent to the subtropics and encompasses Pacific island territories such as Vanuatu, New Caledonia and Fiji. As for $RMM2$, it is often a secondary source of information bearing some added value in the equatorial and eastern part of the domain. The patterns of significant added value for $N34$, $RMM1$ and $RMM2$ are similar to the regression patterns of rainfall against these indices, as illustrated in the first column of Figure 7 and also in several other studies (e.g de Andrade et al., 2018, Figure 2).

## 5.2 Added value in the statistical-dynamical scheme

On account of the decreasing quality of rainfall and MJO forecasts with lead time, it is expected that the importance of the large-scale predictors as illustrated in Section 5.1, but also the importance of the calibration predictor $\hat{r_f}$, might change as we get further away from initialization. This aspect is investigated in Figure 8 with the order of importance of the four predictors at each lead time in the statistical-dynamical approach, when applied to the MF system. Results are similar for the ECMWF system (not shown). Because forecast precipitation was expected to be the n°1 predictor most of the time, we shifted the color code of Figure 7 (second column) from ranks 1-3 to 2-4, and added an extra color for the most important predictor.

At weeks 1 and 2, the forecast precipitation (i.e calibration predictor) is indeed the first source of information in the statistical-dynamical approach almost over the whole domain. In the meantime, the $N34$ predictor often comes second after forecast precipition, at the locations where it is identified as a useful predictor according to Figure 7. The main result is that the prominence of forecast precipitation at these locations shrinks with lead time,

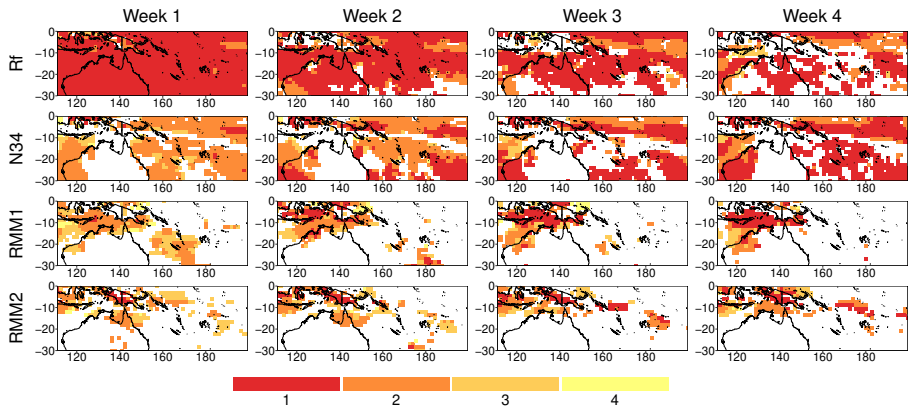


**Fig. 8** Order of selection for each predictor in a stepwise forward selection scheme, applied at each lead time for the calibration + bridging scheme applied to the MF system. Forecast rainfall $\hat{r}_f$ is also included in the selection.

while ENSO gradually becomes the most relevant predictor. As a result, at week 4, $N34$ has become the n°1 predictor for the majority of grid points. This indicates that, as we go into the subseasonal range, the statistical relationship between $N34$ and rainfall actually better captures the ENSO-related predictable signal than the coupled dynamics of the numerical model.

The notable exceptions where $N34$ is challenged by MJO indices are locations where the $RMM1$ predictor or, to a lesser extent, the $RMM2$ predictor, are theoretically good predictors according to Figure 7. This corresponds mostly to the region between the Maritime Continent and northern Australia, where ENSO does not account for much skill. In this area, $RMM1$ is the second most important predictor at weeks 1 and 2 (after precipitation), and it becomes the most important at weeks 3 and 4. As for $RMM2$, it plays a similar role over the land grid points of New Guinea. Figure 7 suggested that these two indices could also be sources of information for wider areas : $RMM1$ can theoretically help predict rainfall in a southeastward-extending region ranging from New Guinea to New Caledonia, and the same can be said of $RMM2$ along the South Pacific Convergence Zone (SPCZ) track. Yet, their added value in these regions at the subseasonal time scales appears to be very limited or non-existent. Admittedly, $RMM1$ comes as second or third predictor in a large zone around New Caledonia at week 1, but this influence shrinks with lead time and ceases to be at weeks 3 and 4. A similar effect can be noted for $RMM2$ over Cape York peninsula in northern Australia.

## 6 Conclusion

We have developed and applied a statistical-dynamical scheme to post-process subseasonal forecasts of precipitation in a bayesian framework. Our framework encompasses a calibration approach that uses the forecast precipitation as a predictor in a statistical model. It also includes bridging aspects under the assumption that large-scale climate features, ENSO and the MJO, are better represented by S2S systems than precipitation itself. Because these large-scale climate features are known to have a strong and direct impact on rainfall on the SWTP domain, we chose this region as a relevant test bed for our scheme.

In this region, the statistical-dynamical approach definitely proves valuable to improve the probabilistic reforecast quality, especially in terms of discrimination for the occurrence vs non-occurrence of a binary event, e.g the upper quintile of weekly precipitation. However, when it comes to deterministic forecasting, the benefits of the approach are more contrasted : depending on the S2S system, it can perform better or worse than a simple bias correction. Therefore, the proposed statistical-dynamical scheme should preferably be applied in a probabilistic forecasting context. These results are based on the Météo-France and ECMWF systems, but were confirmed with four other S2S systems (see Supplementary material, Figures S2 and S3), for which the qualitative benefits are close to those obtained with Météo-France.

The calibration part of the method is the most crucial step as it accounts alone for error reduction, provides reliable probabilistic forecasts, and reasonably improves their discrimination. Yet, adding information from the large-scale predictors is the source of an extra discrimination ability for heavy rainfall events. As a result, although calibration leads to good results on its own, the best of the statistical-dynamical approaches proposed in this study for probabilistic forecasting remains the calibration + bridging approach.

The benefits of using the large-scale predictors actually vary across space and lead times. Unsurprisingly, we have seen that a climate index provides valuable information where the corresponding phenomenon impacts rainfall. To this respect, this article confirms the known impacts of ENSO and the MJO at some locations, e.g the MJO over the Maritime Continent. It also highlights locations that are significantly impacted by the MJO while they do not lie directly within the MJO envelope track, for instance the zone extending southeastward from New Guinea to New Caledonia.

In these locations where MJO-related theoretical predictability is more limited, the degradation of the MJO forecast quality with lead time is such that the MJO predictors do not bring a real predictive improvement of the statistical-dynamical model beyond week 2. The reason why bridging is still profitable at weeks 3 and 4 in these regions is related to the ENSO predictor, that exhibits little variability throughout the forecast, but does not correspond to subseasonal climate variability.

This result suggests that, in the case of SWTP precipitation forecasts, it is difficult to tap into the predictability of large-scale subseasonal variability (the MJO) to obtain improvements in aggregated scores through bridging after week 2. However, it does not exclude the existence of forecasts of opportunity for which particular precipitation events might be predictable three weeks in advance or more (e.g Doss-Gollin et al., 2018; Lin et al., 2019), thanks to subseasonal signals like the MJO. Moreover, it also indicates that other slowly-varying sources of seasonal predictability, analogous to ENSO, could be used to improve subseasonal forecasts. These assumptions could now be tested with similar statistical-dynamical approaches by considering other locations, predictors and predicted variables. For example, the Boreal Summer Intraseasonal Oscillation Index could be tested as a subseasonal predictor for precipitation in the Asian summer monsoon region (Lee et al., 2013), while dynamical weather types could be tested to improve subseasonal prediction of temperatures in Europe (Cassou et al., 2005; Ardilouze et al., 2017).

Finally, some findings in this study point that, beyond statistical post-processing techniques, the improvement of numerical prediction systems is still necessary for the advancement of S2S prediction. This is illustrated by the fact our method does not compensate for the initial difference in skill between the Météo-France and ECMWF systems, while it would if post-processing alone were sufficient to reach the theoretical skill limit. Nonetheless, post-processing proves valuable to diagnose the strengths and weaknesses of S2S systems. For instance, by revealing that a dynamically-forecast large-scale predictor (ENSO) can be more

informative than dynamically-forecast precipitation iteself, it shows that higher skill could actually be obtained if prediction systems directly took advantage of their well-predicted signals. This advocates for a joint development of the S2S systems and their statistical post-processing approaches.

**Acknowledgements** The S2S reforecasts were accessed through the ECMWF MARS data portal (http://apps.ecmwf.int/datasets/data/s2s/). The MSWEP precipitation data is developed by H. Beck (Beck et al., 2017) at Princeton University and is available at http://www.gloh2o.org/. The EOFs used for computation of the RMM indices are provided by the Australian Bureau of Meteorology at http://poama.bom.gov.au/project/maproom/RMM/.

All calculations were performed using the open source language 'R'. The functions in the package 's2dverification' (Manubens et al., 2018) were used for data manipulation. All calculations related to the ROC curve were made with the 'pROC' package (Robin et al., 2011). Empirical CDF determination was made using the 'EnvStats' package (Millard, 2013). Likelihood-ratio tests were performed with the 'lmtest' package.

We thank the two anonymous reviewers whose comments helped improve the former version of this article.

## References

de Andrade FM, Coelho CAS, Cavalcanti IFA (2018) Global precipitation hindcast quality assessment of the Subseasonal to Seasonal (S2S) prediction project models. Climate Dynamics DOI 10.1007/s00382-018-4457-z

Ardilouze C, Batté L, Déqué M (2017) Subseasonal-to-seasonal (S2s) forecasts with CNRM-CM: a case study on the July 2015 West-European heat wave. Advances in Science and Research 14:115–121, DOI 10.5194/asr-14-115-2017

Baggett CF, Nardi KM, Childs SJ, Zito SN, Barnes EA, Maloney ED (2018) Skillful Subseasonal Forecasts of Weekly Tornado and Hail Activity Using the Madden-Julian Oscillation. Journal of Geophysical Research: Atmospheres 123(22), DOI 10.1029/2018JD029059

Beck HE, van Dijk AIJM, Levizzani V, Schellekens J, Miralles DG, Martens B, de Roo A (2017) MSWEP: 3-hourly 0.25° global gridded precipitation (1979-2015) by merging gauge, satellite, and reanalysis data. Hydrology and Earth System Sciences 21(1):589–615, DOI 10.5194/hess-21-589-2017

Bombardi RJ, Pegion KV, Kinter JL, Cash BA, Adams JM (2017) Sub-seasonal Predictability of the Onset and Demise of the Rainy Season over Monsoonal Regions. Frontiers in Earth Science 5, DOI 10.3389/feart.2017.00014

Bröcker J (2015) Resolution and discrimination-two sides of the same coin: Resolution and Discrimination. Quarterly Journal of the Royal Meteorological Society 141(689):1277–1282, DOI 10.1002/qj.2434

Buse A (1982) The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note. The American Statistician 36(3):153–157, DOI 10.2307/2683166

Cassou C, Terray L, Phillips AS (2005) Tropical Atlantic Influence on European Heat Waves. Journal of Climate 18(15):2805–2811, DOI 10.1175/JCLI3506.1

Coelho CA, Pezzulli S, Balmaseda M, Doblas-Reyes F, Stephenson D (2004) Forecast Calibration and Combination: A Simple Bayesian Approach for ENSO. Journal of Climate 17(7):1504–1515

Coelho CA, Firpo MA, de Andrade FM (2018) A verification framework for South American sub-seasonal precipitation predictions. Meteorologische Zeitschrift 27(6):503–520, DOI 10.1127/metz/2018/0898

Cohen J, Coumou D, Hwang J, Mackey L, Orenstein P, Totz S, Tziperman E (2019) S2s reboot: An argument for greater inclusion of machine learning in subseasonal to seasonal forecasts. Wiley Interdisciplinary Reviews: Climate Change 10(2), DOI 10.1002/wcc.567

Dee DP, Uppala SM, Simmons AJ, Berrisford P, Poli P, Kobayashi S, et al. (2011) The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Quarterly Journal of the Royal Meteorological Society 137(656):553–597, DOI 10.1002/qj.828

DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics 44(3):837–845, DOI 10.2307/2531595

Doss-Gollin J, Muñoz AG, Mason SJ, Pastén M (2018) Heavy Rainfall in Paraguay during the 2015/16 Austral Summer: Causes and Subseasonal-to-Seasonal Predictive Skill. Journal of Climate 31(17):6669–6685, DOI 10.1175/JCLI-D-17-0805.1

ECMWF (2017) MJO Index Computation at ECMWF. Tech. rep., ECMWF, URL ftp://acquisition.ecmwf.int/RMMS/RMMS_computation.docx

Gottschalck J, Wheeler M, Weickmann K, Vitart F, Savage N, Lin H, et al. (2010) A Framework for Assessing Operational Madden–Julian Oscillation Forecasts: A CLIVAR MJO Working Group Project. Bulletin of the American Meteorological Society 91(9):1247–1258, DOI 10.1175/2010BAMS2816.1

Hudson D, Alves O, Hendon HH, Marshall AG (2011) Bridging the gap between weather and seasonal forecasting: intraseasonal forecasting for Australia. Quarterly Journal of the Royal Meteorological Society 137(656):673–689, DOI 10.1002/qj.769

Johnson NC, Collins DC, Feldstein SB, L'Heureux ML, Riddle EE (2014) Skillful Wintertime North American Temperature Forecasts out to 4 Weeks Based on the State of ENSO and the MJO*. Weather and Forecasting 29(1):23–38, DOI 10.1175/WAF-D-13-00102.1

Karpechko AY, Charlton-Perez A, Balmaseda M, Tyrrell N, Vitart F (2018) Predicting Sudden Stratospheric Warming 2018 and Its Climate Impacts With a Multimodel Ensemble. Geophysical Research Letters 45(24):13,538–13,546, DOI 10.1029/2018GL081091

Kim H, Vitart F, Waliser DE (2018) Prediction of the Madden–Julian Oscillation: A Review. Journal of Climate 31(23):9425–9443, DOI 10.1175/JCLI-D-18-0210.1

Lee JY, Wang B, Wheeler MC, Fu X, Waliser DE, Kang IS (2013) Real-time multivariate indices for the boreal summer intraseasonal oscillation over the Asian summer monsoon region. Climate Dynamics 40(1-2):493–509, DOI 10.1007/s00382-012-1544-4

Li W, Chen J, Li L, Chen H, Liu B, Xu CY, Li X (2019) Evaluation and Bias Correction of S2s Precipitation for Hydrological Extremes. Journal of Hydrometeorology 20(9):1887–1906, DOI 10.1175/JHM-D-19-0042.1

Liang P, Lin H (2018) Sub-seasonal prediction over East Asia during boreal summer using the ECCC monthly forecasting system. Climate Dynamics 50(3-4):1007–1022, DOI 10.1007/s00382-017-3658-1

Lim Y, Son SW, Marshall AG, Hendon HH, Seo KH (2019) Influence of the QBO on MJO prediction skill in the subseasonal-to-seasonal prediction models. Climate Dynamics 53(3-4):1681–1695, DOI 10.1007/s00382-019-04719-y

Lin H, Mo R, Vitart F, Stan C (2019) Eastern Canada Flooding 2017 and its Subseasonal Predictions. Atmosphere-Ocean 57(3):195–207, DOI 10.1080/07055900.2018.1547679

Luo L, Wood EF, Pan M (2007) Bayesian merging of multiple climate model forecasts for seasonal hydrological predictions. Journal of Geophysical Research: Atmospheres 112(D10), DOI 10.1029/2006JD007655, URL http://doi.wiley.com/10.1029/2006JD007655

Manubens N, Caron LP, Hunter A, Bellprat O, Exarchou E, Fučkar NS, et al. (2018) An R package for climate forecast verification. Environmental Modelling & Software 103:29–42, DOI 10.1016/j.envsoft.2018.01.018

Marshall AG, Hendon HH (2015) Subseasonal prediction of Australian summer monsoon anomalies. Geophysical Research Letters 42(24):10913–10919, DOI 10.1002/2015GL067086

Marshall AG, Hendon HH (2019) Multi-week prediction of the Madden–Julian oscillation with ACCESS-S1. Climate Dynamics 52(5-6):2513–2528, DOI 10.1007/s00382-018-4272-6

Marshall AG, Hudson D, Wheeler MC, Hendon HH, Alves O (2011) Assessing the simulation and prediction of rainfall associated with the MJO in the POAMA seasonal forecast system. Climate Dynamics 37(11-12):2129–2141, DOI 10.1007/s00382-010-0948-2

Mastrangelo D, Malguzzi P (2019) Verification of Two Years of CNR-ISAC Subseasonal Forecasts. Weather and Forecasting 34(2):331–344, DOI 10.1175/WAF-D-18-0091.1

McGree S, Whan K, Jones D, Alexander LV, Imielska A, Diamond H, et al. (2014) An updated assessment of trends and variability in total and extreme rainfall in the western Pacific. International Journal of Climatology 34(8):2775–2791, DOI 10.1002/joc.3874

McGree S, Schreider S, Kuleshov Y (2016) Trends and Variability in Droughts in the Pacific Islands and Northeast Australia. Journal of Climate 29(23):8377–8397, DOI 10.1175/JCLI-D-16-0332.1

Millard SP (2013) EnvStats. Springer New York, DOI 10.1007/978-1-4614-8456-1

Murphy AH (1973) A New Vector Partition of the Probability Score. Journal of Applied Meteorology 12(4):595–600

Pegion K, Kirtman BP, Becker E, Collins DC, LaJoie E, Burgman R, et al. (2019) The Subseasonal Experiment (SubX): A multi-model subseasonal prediction experiment. Bulletin of the American Meteorological Society DOI 10.1175/BAMS-D-18-0270.1

Robertson AW, Vitart F (2019) Sub-seasonal to seasonal prediction: the gap between weather and climate forecasting. Elsevier

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12(1):77, DOI 10.1186/1471-2105-12-77

Schepen A, Wang QJ, Robertson DE (2012) Combining the strengths of statistical and dynamical modeling approaches for forecasting Australian seasonal rainfall. Journal of Geophysical Research: Atmospheres 117(D20), DOI 10.1029/2012JD018011

Schepen A, Wang QJ, Robertson DE (2014) Seasonal Forecasts of Australian Rainfall through Calibration and Bridging of Coupled GCM Outputs. Monthly Weather Review 142(5):1758–1770, DOI 10.1175/MWR-D-13-00248.1

Schepen A, Wang QJ, Everingham Y (2016) Calibration, Bridging, and Merging to Improve GCM Seasonal Temperature Forecasts in Australia. Monthly Weather Review 144(6):2421–2441, DOI 10.1175/MWR-D-15-0384.1

Specq D, Batté L, Déqué M, Ardilouze C (2020, in press) Multimodel forecasting of precipitation at subseasonal timescales over the southwest tropical Pacific. Earth and Space Science DOI 10.1029/2019EA001003

Strazzo S, Collins DC, Schepen A, Wang QJ, Becker E, Jia L (2019) Application of a Hybrid Statistical–Dynamical System to Seasonal Prediction of North American Temperature and Precipitation. Monthly Weather Review 147(2):607–625, DOI 10.1175/MWR-D-18-0156.1

Tian D, Wood EF, Yuan X (2017) CFSv2-based sub-seasonal precipitation and temperature forecast skill over the contiguous United States. Hydrology and Earth System Sciences

21:1477–1490, DOI 10.5194/hess-21-1477-2017

Vigaud N, Robertson AW, Tippett MK (2017a) Multimodel Ensembling of Subseasonal Precipitation Forecasts over North America. Monthly Weather Review 145(10):3913–3928, DOI 10.1175/MWR-D-17-0092.1

Vigaud N, Robertson AW, Tippett MK, Acharya N (2017b) Subseasonal Predictability of Boreal Summer Monsoon Rainfall from Ensemble Forecasts. Frontiers in Environmental Science 5:67, DOI 10.3389/fenvs.2017.00067

Vigaud N, Tippett MK, Robertson AW (2018) Probabilistic Skill of Subseasonal Precipitation Forecasts for the East Africa–West Asia Sector during September–May. Weather and Forecasting 33(6):1513–1532, DOI 10.1175/WAF-D-18-0074.1

Vigaud N, Tippett MK, Yuan J, Robertson AW, Acharya N (2019) Spatial correction of multi-model ensemble subseasonal precipitation forecasts over North America using local Laplacian eigenfunctions. Monthly Weather Review DOI 10.1175/MWR-D-19-0134.1

Vitart F (2004) Monthly Forecasting at ECMWF. Monthly Weather Review 132(12):2761–2779, DOI 10.1175/MWR2826.1

Vitart F (2014) Evolution of ECMWF sub-seasonal forecast skill scores. Quarterly Journal of the Royal Meteorological Society 140(683):1889–1899, DOI 10.1002/qj.2256

Vitart F (2017) Madden-Julian Oscillation prediction and teleconnections in the S2s database. Quarterly Journal of the Royal Meteorological Society 143(706):2210–2220, DOI 10.1002/qj.3079

Vitart F, Ardilouze C, Bonet A, Brookshaw A, Chen M, Codorean C, et al. (2017) The Subseasonal to Seasonal (S2S) Prediction Project Database. Bulletin of the American Meteorological Society 98(1):163–173, DOI 10.1175/BAMS-D-16-0017.1

Wang L, Robertson AW (2019) Week 3–4 predictability over the United States assessed from two operational ensemble prediction systems. Climate Dynamics 52(9-10):5861–5875, DOI 10.1007/s00382-018-4484-9

Wheeler MC, Hendon HH (2004) An All-Season Real-Time Multivariate MJO Index: Development of an Index for Monitoring and Prediction. Monthly Weather Review 132(8):1917–1932, DOI 10.1175/1520-0493(2004)132¡1917:AARMMI¿2.0.CO;2

White CJ, Carlsen H, Robertson AW, Klein RJ, Lazo JK, Kumar A, et al. (2017) Potential applications of subseasonal-to-seasonal (S2S) predictions. Meteorological Applications 24(3):315–325, DOI 10.1002/met.1654

Wilks DS (2006) Statistical methods in the atmospheric sciences, 2nd edn. Academic Press, Boston, MA

Zhang C (2005) Madden-Julian Oscillation. Reviews of Geophysics 43(2), DOI 10.1029/2004RG000158

Zhang C (2013) Madden-Julian Oscillation: Bridging Weather and Climate. Bulletin of the American Meteorological Society 94(12):1849–1870, DOI 10.1175/BAMS-D-12-00026.1

## Conflict of interest

The authors declare that they have no conflict of interest.