Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Home Affairs FDHA
Federal Office of Meteorology and Climatology MeteoSwiss

MeteoSwiss

Technical Report MeteoSwiss No. 263

Systematic quality assessment of an operational seasonal forecasting system

Kathrin Wehrli, Jonas Bhend, and Mark A. Liniger

ISSN: 2296-0058

Systematic quality assessment of an operational seasonal forecasting system

Kathrin Wehrli, Jonas Bhend, and Mark A. Liniger

Recommended citation:

Wehrli, K., J. Bhend and M. A. Liniger: 2017, Systematic quality assessment of an operational seasonal forecasting system *Technical Report MeteoSwiss,* **263,** 52 pp.

Editor:

Federal Office of Meteorology and Climatology, MeteoSwiss, © 2017

MeteoSwiss
Operation Center 1
CH-8044 Zürich-Flughafen
T +41 58 460 99 99
www.meteoschweiz.ch

Abstract

Seasonal forecasts are increasingly taken into account in the decision processes of various sectors. The predictability of the climate several months ahead varies strongly in space and time. Detailed information on the quality of the forecasts is compulsory to make optimal use of the available forecasts. Such information, however, is not readily available. We have compiled a comprehensive validation of forecasts by an operational seasonal forecasting system, the ECMWF System 4. The forecasts for temperature and precipitation are verified against the ERA-Interim reanalysis, a proxy for observations, using a set of skill metrics describing various aspects of forecast quality. This report documents the data and methodology used and illustrates the variability of seasonal forecast skill by presenting results for selected seasons and regions of the world. To explore the space-time variability of forecast skill and to engage with users, we have developed an interactive web-app. This low-threshold tool allows stakeholders to assess the quality of seasonal forecasts specifically for their region, season, forecast horizon and parameter of interest.

Zusammenfassung

Saisonale Klimavorhersagen werden in verschiedenen Sektoren immer häufiger beim Entscheidungsfindungsprozess herangezogen. Die Vorhersagbarkeit des Klimas über mehrere Monate hinweg schwankt stark im Raum und mit der Zeit. Um die Vorhersagen bestmöglich zu nutzen, braucht es Angaben zur Qualität der Vorhersagen. Ein Katalog, der diese Informationen für die operationellen Vorhersagesysteme bereitstellt, ist bis zum jetzigen Zeitpunkt jedoch nicht verfügbar. Wir haben eine umfassende Verifikation von Vorhersagen für ein operationelles saisonales Vorhersagesystem, ECMWF System 4, erstellt. Die Temperatur- und Niederschlags-Vorhersagen werden verifiziert gegen die ERA-Interim Reanalysen, welche als Proxy für Beobachtungen dienen. Die Güte der Prognosen wird mithilfe einer Reihe von Gütemassen bewertet, welche verschiedene Aspekte der Vorhersagequalität abdecken. Dieser Fachbericht dokumentiert die verwendeten Datensätze und Methoden zur Verifikation. Die Variabilität der Güte von saisonalen Vorhersagen wird illustriert anhand von Resultaten für verschiedene Regionen der Erde. Um anderen zu ermöglichen die Raum-Zeit Variabilität der Vorhersagegüte selber zu erforschen, haben wir zudem eine interaktive Webapplikation entwickelt. Mit Hilfe dieses Tools können Akteure verschiedener Sektoren die Qualität saisonaler Vorhersagen spezifisch für ihre Region untersuchen und finden Informationen für die gewünschte Saison, den relevanten Vorhersagehorizont und den Parameter von Interesse.

Contents

1 Introduction

Operational seasonal forecasts are produced now by various forecasting centers. Steadily increasing computer power, better understanding of the climate system, the development of coupled atmosphere-ocean models and the introduction of ensemble approaches made it possible to produce skillful forecasts beyond the weather time scale (Buizza and Leutbecher, 2015). In the past, decision-makers had to rely on historical data for their planning and decisions in many sectors such as energy, agriculture, health, water management and tourism. Since the quality of seasonal forecasts is improving, their potential use for decision making is increasing.

The EUPORIAS (short for European Provision Of Regional Impacts Assessments on Seasonal and Decadal Timescales) project aims to provide climate services based on long-range forecasts fitted to specific user needs. One of the key findings of this project is that users have to be provided with information on forecast quality along with the actual forecasts (Taylor et al, 2015). Users need skill information in a systematic way for all forecasts that are provided by the operational forecast producers. This is particularly important due to the fact that the predictability and skill strongly vary from month to month and in space (as was demonstrated for medium-range forecasts by Buizza and Leutbecher (2015). Such a comprehensive skill assessment of the operational models has not been available so far. It is the goal of this work to fill this gap and implement the skill assessment from a user perspective. The results shall support the users in their work with seasonal forecasts beyond the EUPORIAS project and help to transfer knowledge gained in the project into operational practice.

We realize this comprehensive skill assessment for the System 4 model, which is the operational seasonal forecasting system of the European Centre for Medium-range Weather Forecast's (ECMWF). Global monthly and seasonal forecasts for temperature and precipitation are verified for forecasts up to 7 months in advance. A systematic skill analysis of the ECMWF System 4 model is presented for all starting dates and all lead times using a range of verification metrics. In section 2 of this report the data used and verification methodology is described and the verification metrics are explained. In section 3 results from the skill analysis are presented for the temperature and precipitation forecast based on examples of different regions. To provide the users an opportunity to explore the space-time variability of forecast skill, the results are made publicly available via an interactive web application, which is also documented in section 4 of this report. The final section presents the conclusions and an outlook.

2 Data and Methods

2.1 Seasonal forecast data and verifying observations

We analyze monthly temperature and precipitation forecasts from the ECMWF System 4 forecasting system. ECMWF System 4 is a coupled ocean-atmosphere model that provides seasonal forecasts in real-time and forecasts back to 1981 (Molteni et al., 2011). Hindcasts (also to be referred to as re-forecasts) cover the years 1981 to 2010 before the model was ready to produce operational forecasts in 2011. Here we analyze the hindcasts and operational forecasts from 1981 throughout 2014, and we will refer to them generally as 'forecasts' if no distinction has to be made. The forecasts are initialized at the 1st of each month and are run for the following 7 months. Here we say that the forecasts have lead times 1 to 7 months. The term lead time refers to the period of time between when the forecast is issued and for when it is valid. A forecast that is issued for the same month is said to have 1 month lead time, since it is based on data up to the 1st of that month and is valid for the upcoming month. For example, a forecast for December issued in November is said to have a lead time of 2 months.

ECMWF System 4 produces global forecast data with a grid spacing of about 0.7 degrees (Molteni et al., 2011). The operational version of ECMWF System 4 runs with 51 ensemble members while the hindcasts have been run with 15 members (except for hindcasts initialized in February, May, August, and November, for which also 51 members have been run). To have a consistent dataset we decided to consider 15 members for the whole analysis (the first 15 members were taken). For computational reasons, we have regridded the forecast data to a 2° grid using bilinear interpolation prior to the skill analysis. This was done in order to save computational time during the skill analysis.

The forecasts are verified against ERA-Interim (Dee et al., 2011), which is the latest continuously updated reanalysis dataset produced by ECMWF. The ERA-Interim reanalysis assimilates numerous observations to a fixed dynamical model version to produce a physically consistent dataset containing the global best estimate of ocean, land-surface, weather and upper-air parameters. ERA-Interim has the same horizontal resolution as ECMWF System 4, which is 0.7 degrees or about 79 km. Previous to any further analyses; we have interpolated the reanalysis to the same grid as the forecasts for use in this study.

The rationale behind using a reanalysis dataset for skill verification instead of e.g. gridded observational data is its representativeness. Especially in model studies, the spatial and temporal completeness is a major advantage. Similar spatial (and also temporal) scales are represented in the reanalysis as in the forecast models, which allows direct comparison. Cornes and Jones (2014) have assessed temperature trends in Europe for ERA-Interim and show that the reanalysis is not only good at replicating trends in means but is also reliable for trends in temperature extremes. Reliability of the

reanalysis depends on the data assimilation scheme and the numerical model. The reanalysis uses a forecast model to assimilate the observations from multiple sources and with that ensures that the estimated parameters are physically consistent. Beside the model formulation, a key limitation of reanalyses is connected to inconsistencies of the underlying observational data (Dee, 2012). The reliability can vary substantially between certain regions, time periods and the variable considered. We will have to keep this in mind when interpreting the results.

It must be kept in mind, that ERA-Interim and System 4 share the same atmospheric dynamical model (of a different version though). Errors of the model could be common both for the reanalysis and the seasonal forecasts. Therefore, the presented assessment must be assumed to be an upper estimate of predictive skill. Generally speaking, seasonal forecasts compared to station based observations result in lower skill estimates.

2.2 Methodology

Forecast quality is multifaceted. No single verification metric exists that captures all aspects of forecast quality. Therefore, we verify the ECMWF System 4 forecasts using a range of verification metrics introduced below.

Some of these metrics are sensitive to systematic biases in the forecasts. We correct systematic offsets in monthly and seasonal means using a simple mean bias correction introduced in the following section.

2.2.1 Bias correction

We apply an additive monthly and seasonal mean bias correction to the temperature forecasts, and a multiplicative bias correction (scaling) to the precipitation forecasts. To mimic real forecast situations and assess how well the prediction system works when the bias correction parameters are not known, we apply the bias correction in a leave-one-out cross-calibration mode. This is achieved by separating the actual dataset of observations and forecasts into two subsets, the first of which is used to estimate the bias correction parameters (i.e. the systematic model error) with which the second subset is calibrated. To mimic the calibration for an operational, future forecast and to minimize sampling errors in the parameter estimates, we calibrate one year at a time using the 33 remaining years in the series to estimate the bias. By iteratively cycling through all available years, we obtain the bias-corrected series of 34 years of data.

2.2.2 Verification metrics

We measure different attributes of forecast quality (Murphy, 1993) using a set of bespoke verification metrics. These metrics are introduced later in this section. The aspects of forecast quality covered are the following:

- **Association:** How well do variations in the forecast ensemble mean correspond to variations in the observations? We use the correlation coefficient to measure the strength of this relationship in a linear way.

- **Accuracy:** How well does the forecast predict the category that the observation falls into and how close is the continuous ensemble forecast to the observations? This quality describes the level of agreement between forecasts and observations and we used the Fair Ranked Probability Skill Score and Fair Continuous Ranked Probability Skill Score as measure.
- **Reliability:** How well do the forecast probabilities match the observed frequencies? We measure this quality using the Fair Spread to Error Ratio.
- **Discrimination:** How well can the forecast successfully distinguish outcomes for which the observations differ? The ability of the forecast to discriminate among observations is measured using the Generalized Discrimination Score for ensemble forecasts and the ROC area score for probabilistic multi-category forecasts.

Where possible, the 'fair' version of the verification metrics is used as for example with the Fair Continuous Ranked Probability Skill Score. Accuracy measures such as the Ranked Probability Skill Score have been shown to systematically deteriorate for small ensembles (Müller et al., 2005a; Weigel et al., 2007; Weigel, 2012). Fair scores are used to mitigate the systematic effects of limited ensemble size on these scores (Ferro et al., 2008; Ferro, 2014). The fair scores represent more realistically the information available in a real-time forecast. The operational forecast system uses 51 ensemble members while in this skill assessment 15 members are used. Using the fair scores that reflect the expected score for an infinite ensemble, the verification metrics are more representative of the expected skill for a future, operational forecast with 51 members.

The bias-corrected forecasts described earlier in section 2.2.1 are used to compute the Continuous Ranked Probability and the Spread to Error skill scores. It is not necessary to apply a bias correction to the other scores since they are not sensitive to forecast bias.

Correlation

The strength of a linear relationship between forecasts and observations is measured using the correlation coefficient. We use the Pearson correlation coefficient to compute the association between the ensemble mean and the reanalysis data. This measure does not take forecast bias into account but it is sensitive to outliers. The correlation ranges from -1 to 1. A forecast with perfect association has a correlation of 1 and if the coefficient is -1 the ensemble means and observations are perfectly anti-correlated. Values close to 0 indicate weak correlation. A constant climatological forecast exhibits a correlation of 0.

For testing significance of the results, the 95% confidence intervals of a one-sided Student's t-test are computed. In the graphics, stippling of the respective grid cell indicates correlation significantly larger than zero at the 5% level of significance. With the 34 years of hindcasts available for verification, correlations > 0.3 indicate significantly positive association (at the 5% level). In general, the forecasts are not independent due to the presence of trends and long-range memory (autocorrelation). Dependence of forecasts increases the sampling error of correlation estimates. We have not attempted to quantify the effect of dependence on significance levels here and we therefore recommend treating the significance assessment with caution.

Ranked Probability Skill Score

The Ranked Probability Score (RPS) is a measure of the accuracy of probabilistic multi-category forecasts (Epstein, 1969; Murphy, 1969, 1971). This forecasts and observations are separated in K categories and the verification is carried out based on the statistical distribution across these categories. The RPS is the squared difference between the cumulative distribution function of the forecast (CDF_{fc}) and the observation (CDF_{obs}) as defined in equation (1). As the RPS is a quadratic measure, larger deviations from the actual probability (i.e. forecast probabilities in categories further away from the observed category) are penalized stronger than smaller ones. For a perfect forecast one gets RPS=0 and the score is larger than 0 for imperfect forecasts.

$$RPS = \frac{1}{K-1}\sum_{k=1}^{K}\left(CDF_{fc,k} - CDF_{obs,k}\right)^2 \quad (1)$$

The ranked probability skill score (RPSS) describes the relative accuracy of the forecasting system with respect to a reference forecast. Once the RPS is known the corresponding skill score is calculated from the fraction of the forecast RPS and the reference RPS as in equation (2). Usually the reference forecast is a constant probabilistic climatology forecast or persistence. Here we use a constant probabilistic forecast that is expected to follow the long-term probability of an event. In the case of a tercile forecast, the climatological forecast is one third for each of the three classes.

$$RPSS = 1 - \frac{RPS}{RPS_{clim}} \quad (2)$$

The RPSS answers the question, how much added value a forecast provides with respect to a climatological forecast in predicting the category that the observations fall into. A prediction system has perfect skill when RPSS=1, if RPSS>0 the forecast has some skill over the climatological forecast and there is no skill if the score is equal or smaller than 0. For a large sample, random forecasts result in a value of 0.

Here the skill of the system to forecast a three-category event was assessed. The meaning of the three forecast categories may be stated as:

- above normal: temperature/precipitation in the warmest/wettest one-third when compared to the defined reference period (here the reference period is defined as the preceding years from 1981 to 2014);
- below normal: temperature/precipitation in the coolest/driest one-third when compared to the defined reference period;
- near normal: temperature/precipitation in the middle one-third relative to the defined reference period.

In the graphics, forecasts significantly (at the 5% level) better than climatology are indicated by stippling of the respective grid cells. Significance of the skill scores is estimated using the standard error of the skill score estimated by propagation of error (Siegert, 2015).

Continuous Ranked Probability Skill Score

The Continuous Ranked Probability Score (CRPS) is an extension of the RPS for continuous probability forecasts (Matheson and Winkler, 1976; Hersbach, 2000). It measures the integrated squared

difference between the cumulative distribution function of the forecasts and the observations not across categories, but for the parameter of interest x (e.g. 2m temperature or precipitation):

$$CRPS = \int_{-\infty}^{\infty} [P(x) - P_a(x)]^2 \, dx \qquad (3)$$

Here P(x) is a cumulative distribution over the PDF of a probabilistic forecast (ρ) given by:

$$P(x) = \int_{-\infty}^{x} \rho(y) \, dy$$

Thus P(x) measures the probability the prediction system forecasts for the case that the observed outcome x_a is smaller than x (Hersbach 2000), which is between zero and one hundred percent. $P_a(x)$ is a cumulative distribution given by the Heaviside function H(x):

$$P_a(x) = H(x - x_a), \text{ with } H(x) = \begin{cases} 0 & for\ x < 0 \\ 1 & for\ x \geq 0 \end{cases}$$

For a perfect deterministic forecast P = P_a is fulfilled and the CRPS gets a value of zero. The CRPS is a measure of the accuracy of the forecasts and is sensitive to both the reliability and sharpness of the forecast. That is, forecasts get a lower CRPS if they "mean what they say" (reliability) and if they take a risk and are thus different from the climatological forecast (sharpness).

The Continuous Ranked Probability Skill Score (CRPSS) is defined as the relative accuracy compared with a reference forecasts analogous to the RPSS in equation (2). Here we use a constant probabilistic climatological forecast derived from the observations as the reference. If CRPSS > 0, forecasts are more accurate than the reference; CRPSS = 1 denotes a perfect deterministic (i.e. no ensemble spread) forecast.

The CRPSS is computed using the bias-corrected forecasts. Like for the RPSS, forecasts significantly (at the 5% level) better than climatology are indicated by stippling of the respective grid cells in graphics showing the CRPSS.

Generalized Discrimination Score

The ability of the forecast system to tell cases with different observed outcomes apart is measured using the Generalized Discrimination Score. It is calculated by comparing observation and forecast pairs and answers the question of how often the forecast successfully distinguishes outcomes for which the observations differ. We use the Generalized Discrimination Score for ensemble forecasts as introduced by Weigel and Mason (2011). For each pair of observations in the verification set, the corresponding pair of forecasts is awarded a score of 1 if the forecasts differ the same way the observations differ (i.e. for forecasts y and observations x: if $y_1 > y_2$ and $x_1 > x_2$), 0.5 if the forecasts or observations are identical, or 0 if the forecasts differ inversely from the observations (i.e. $y_1 < y_3$ but $x_1 > x_3$). The Generalized Discrimination Score is then the average of the scores for all possible pairs of observations (forecasts) in the verification set. A Generalized Discrimination Score of 0.5 or 50% means that the forecast has no discrimination; this is what would be achieved by random guessing. A Generalized Discrimination Score of 1 indicates perfect discrimination and a value of 0 means that the forecast is perfectly bad.

The Generalized Discrimination Score for continuous probability forecasts corresponds to the ROC score for a categorical forecast (see below for the ROC score). The Generalized Discrimination Score does not assess the reliability and accuracy of the forecast, as it is not sensitive to bias. Thus it can be considered a measure of potential usefulness.

Spread to Error Ratio

To assess the forecast reliability, we compute the Spread to Error Ratio (SPR). Reliability indicates how well the forecast probabilities match the observed frequency of occurrence. The SPR quantifies the ability of the ensemble forecasts to represent the forecast error in a statistical sense. The SPR relates the time mean ensemble variance (σ_e^2) or spread with the mean squared error (MSE) of the forecast ensemble mean (Weigel, 2012, Ho et al., 2013). The condition given in equation (4) for m ensemble members (here m is 15) must be fulfilled for a forecast to be considered reliable.

$$\sigma_{\bar{e}}^2(\tau) = \frac{m}{m+1} MSE(\tau) \quad (4)$$

Therefore, to assess forecast reliability by the Spread to Error Ratio we consider the ratio of σ_e^2 to the Root Mean Squared Error of the forecast ensemble mean and adjust by the ensemble size factor in equation (4) resulting in equation (5).

$$SPR = \sqrt{\frac{m+1}{m}} \frac{\sigma_e}{RMSE} \quad (5)$$

A SPR of 1 is a necessary but not sufficient condition for forecasts to be reliable. SPR < 1 indicates overconfidence whereby the ensemble spread underestimates the forecast uncertainty. SPR > 1 indicates overdispersion. On seasonal timescales, most forecasting systems tend to be overconfident.

The Spread to Error Ratio is computed using the bias-corrected forecasts.

ROC area score

The area under the ROC curve (Receiver Operating Characteristic curve, Peterson and Birdsall, 1953; for application in atmospheric science e.g. Mason, 1982, Harvey et al., 1992; Mason and Graham, 1999) is a measure of discrimination for dichotomous forecasts (occurrence and non-occurrence of an event). The ROC curve is created by plotting the hit rate on the y-axis versus the false alarm rate for a set of increasing probability thresholds. The hit rate is the fraction of the events that was forecasted correctly, the false alarm rate indicates how often an event was forecasted but did not occur. How to calculate the hit rate and false alarm rate can be derived from the contingency table shown in Table 1 for the forecast of a dichotomous event. Usually one wants to have probabilistic forecasts that do not only predict if an event occurs or not but assign probabilities to events. In this case a 2 x n_f frequency distribution table is created with n_f being the number of probability categories (Harvey et al., 1992). In the case of categorical forecasts, a ROC curve is created for each of the forecast categories using separate contingency tables (three ROC curves in the case of tercile forecasts).

The ROC area score describes the forecast discrimination using the area below the ROC curve. It is a special case of the Generalized Discrimination Score for probabilistic forecasts predicting dichoto-

mous events (Mason and Weigel, 2009). Examples of ROC curves for the three cases are shown in Figure 1. A forecast with perfect discrimination has a ROC score of 1. A forecast that is not better than guessing has a ROC score of 0.5 (corresponds roughly to the plot shown in Figure 1 in the middle) and a perfectly bad forecast has a score of 0.

Table 1: 2 x 2 contingency table to evaluate the forecasting performance of a dichotomous forecast, adapted from Harvey et al. (1992). A dichotomous forecast will either predict that an event occurs or that it will not occur. In the future the event is then observed or it does not occur. Thus a forecasting system with two probability categories has four possible outcomes, which are described in the table. The hit rate and false-alarm rate, which are used to create the ROC curve, are given at the bottom of the table.

	Event forecast	
Event	**Will occur**	**Will not occur**
Occurs	A (hit)	C (miss)
Does not occur	B (false alarm)	D (correct rejection)
Probability of detection = hit rate =	$\frac{A}{A+C}$	
False-alarm rate =	$\frac{B}{B+D}$	

The ROC score does not assess the reliability or accuracy of the forecast since it is not sensitive to bias in the forecast. Thus it can be considered a measure of potential usefulness. Here the discrimination of the system to forecast a three-category event was assessed. The three categories are defined like for the RPSS described earlier in this chapter. The area under the ROC curve is compared to a reference forecast (climatology), which gives us the skill score for each forecast tercile. In the following we show only ROC scores for upper and lower tercile categories. In the graphics, stippling of the respective grid cells indicates where forecasts are significantly (at the 5% level) better than guessing.

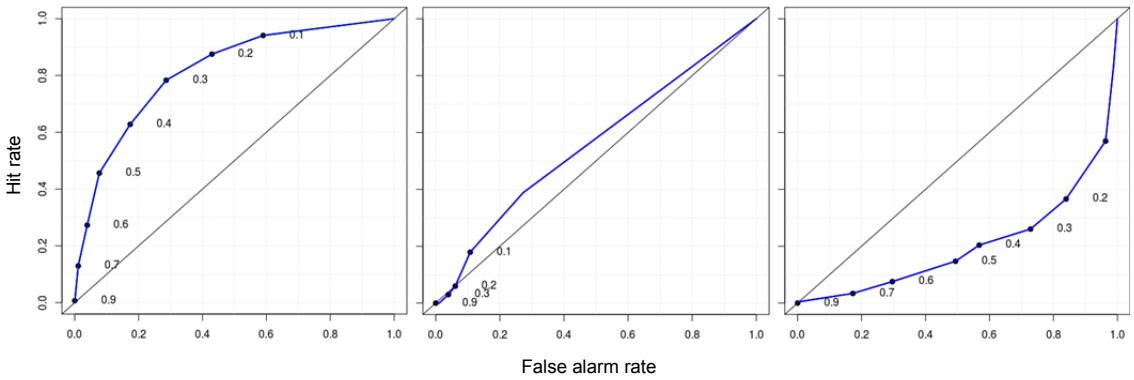

Figure 1: Three examples for Relative Operating Characteristic (ROC) curves. ROC curves are created by plotting the hit rate against the false alarm rate for a set of increasing probability thresholds (marked by black dots and labelled with decimals between 0 and 1 in the figure). These plots were created from artificially generated data. The black line from the lower left to the upper right corner is the 45-degree diagonal of the ROC space and denotes a test with no discrimination. The curve of a perfect forecast would run from the lower left corner over the upper left to the upper right corner. The plot to the left shows a forecast with a ROC score between 0.5 and 1, hence this forecast has discrimination (but is not perfect). The plot in the middle shows a forecast with a ROC score close to 0.5, which means that the forecast has nearly no skill. The plot to the right shows a forecast that performs even worse than guessing and has a ROC score between 0 and 0.5.

3 Forecast quality verification results

In this chapter exemplary results from the skill analysis are presented, which show well-known features of seasonal climate variability but also point to unusual patterns in forecast skill. The relation between the skill scores is illustrated with the help of cases from the skill assessment. Focus is put on selected regions of the world to examine the temporal and spatial variability of forecast skill.

3.1 Forecast skill in the El Niño region

The El Niño Southern Oscillation (ENSO) is a climate variability phenomenon coupling the atmospheric circulation and the ocean in the tropical eastern Pacific Ocean. Its impacts on weather, climate and as a consequence on society and economy are nearly global. ENSO oscillates between a warm phase, called El Niño and a cool phase, La Niña. The two phases typically last for several months and occur with a periodicity of three to seven years (McPhaden, 2003). Extreme climate conditions like droughts, heat waves and floods can be associated with ENSO events. Forecasting of the ENSO mode supplies valuable information for local agriculture, fishery, health and safety but also for users worldwide. The intensity of an ENSO event is classified using anomalies of the sea-surface temperature (SST) in a pre-defined region in the Pacific. The most commonly used region is Niño3.4, which is defined as the area between 5 °N to 5 °S and 170 °W to 120 °W (see Figure 3). Positive anomalies are recorded in case of an El Niño and negative anomalies in case of a La Niña event.

The interannual variability of the SST in the tropical Pacific associated with ENSO is one of the main sources of predictability of seasonal mean variability. The global response to the anomaly in tropical heating contributes significantly to the skill of seasonal forecasts (Shukla et al., 2000; Hoerling and Kumar, 2002; Jha et al., 2016). A map of the most common impacts related to ENSO teleconnections is shown in Figure 2 for the winter season (December to February), when ENSO is strongest. Mainly the same areas are affected by El Niño and La Niña – impacts in one of the phases are usually opposite to the other. Direct impacts are commonly experienced all over the world, except for Europe, large parts of Asia and the polar regions.

Figure 3 shows the Correlation coefficient of temperature (at the top) and precipitation (at the bottom) for the November-December-January (NDJ) season from forecasts initialized in October, which means that lead time is 2 to 4 months. The most prominent feature in global forecast skill is the outstanding forecast performance in the tropical Pacific, due to the ENSO phenomenon. One can clearly identify the ENSO region by the high correlation of the temperature forecasts with the observations of over 90%. The effect of ENSO extends outside the Niño3.4 region (black rectangle in Figure 3). We will refer to the ENSO region in more general terms as an area in the tropical Pacific where forecasts are highly influenced by ENSO phase, which is encompassing but not restricted to the Niño3.4 region.

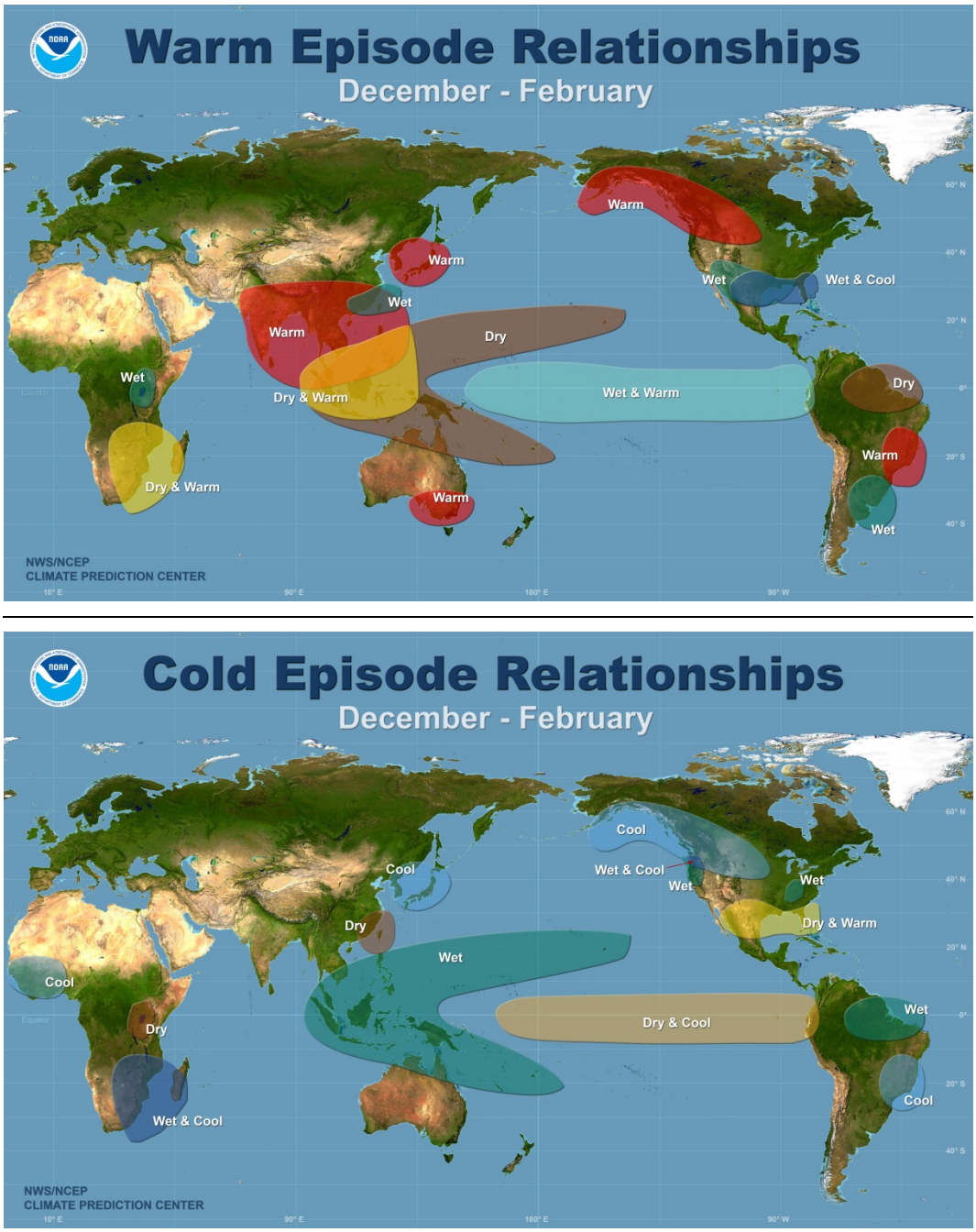

Figure 2: The most common impacts on temperature and precipitation related to ENSO for the peak season of the El Niño or La Niña phase, during December to February. The El Niño teleconnections ("warm episode") are shown at the top and La Niña ("cold episode") at the bottom. Note that impacts of ENSO are also experienced during the rest of the year but not shown in these graphics for simplicity. Both images courtesy NWS/NCEP Climate Prediction Center, retrieved from https://www2.ucar.edu/.

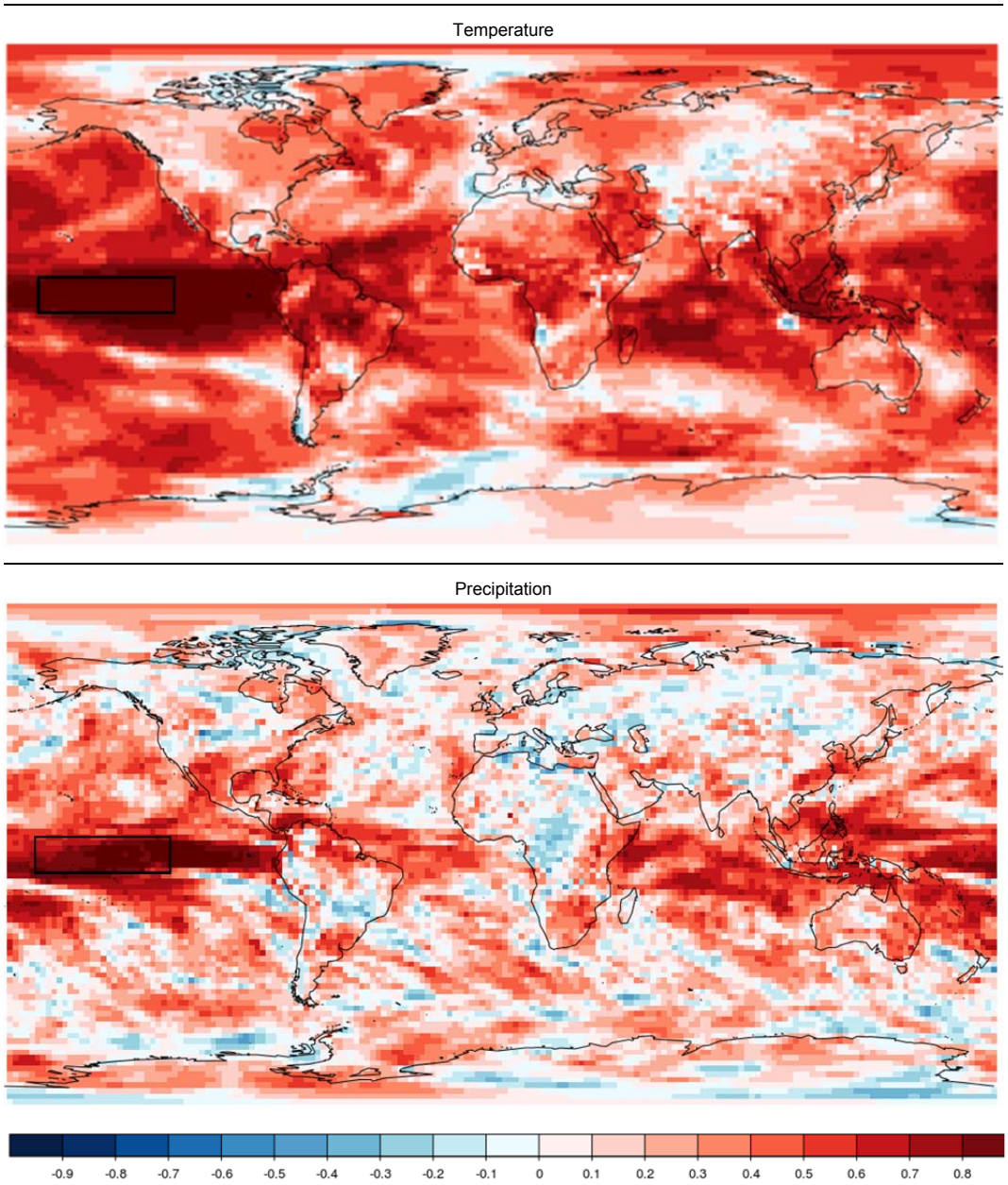

Figure 3: Correlation of the November-January (NDJ) temperature (top) and precipitation (bottom) forecasts initialized in October (lead time of the forecast is 2 to 4 months). Darker colors indicate stronger linear relationship between the forecasts and the observations. The black rectangle marks the location of the Niño3.4 region. Stippling for significantly positive correlations is not shown for clarity. Correlations exceeding 0.3 are significantly (at the 5% level) larger than zero.

The ENSO signal is also seen in the precipitation forecast skill due to the short lead time chosen. Furthermore, regional patterns of increased forecast skill are recognizable that can be attributed to the impacts of ENSO. Comparing Figure 3 with Figure 2, some of the areas identified are for example Indonesia and the Indian Ocean, Eastern and South Africa, Eastern Australia, Mexico, Northern Brazil and the mainland around the ocean bay between Uruguay and Argentina. In these areas forecasts seem to perform particularly well due to the predictability of ENSO teleconnections.

In the following, we examine the variability with lead time and target season for the tropical Pacific, in particular the ENSO region. Figure 4 shows the Fair RPSS over the tropical eastern Pacific for tem-

perature forecasts with lead times [2,3,4] at the bottom and [5,6,7] at the top and for two target seasons (NDJ to the left and MJJ to the right). Grid cells where forecasts perform significantly better than guessing outcomes from climatology are indicated by stippling. This applies to the ENSO region, which clearly stands out in the plots for NDJ. The skill of the forecasts increases steadily when going to shorter lead times (from top to bottom in Figure 4). An increase of skill when coming closer to the target season is also shown for the MJJ forecasts. However, when comparing with the NDJ season the forecasts are less skillful for all lead times. Furthermore the ENSO region does not appear as distinct as for NDJ and the forecast skill distribution looks more irregular.

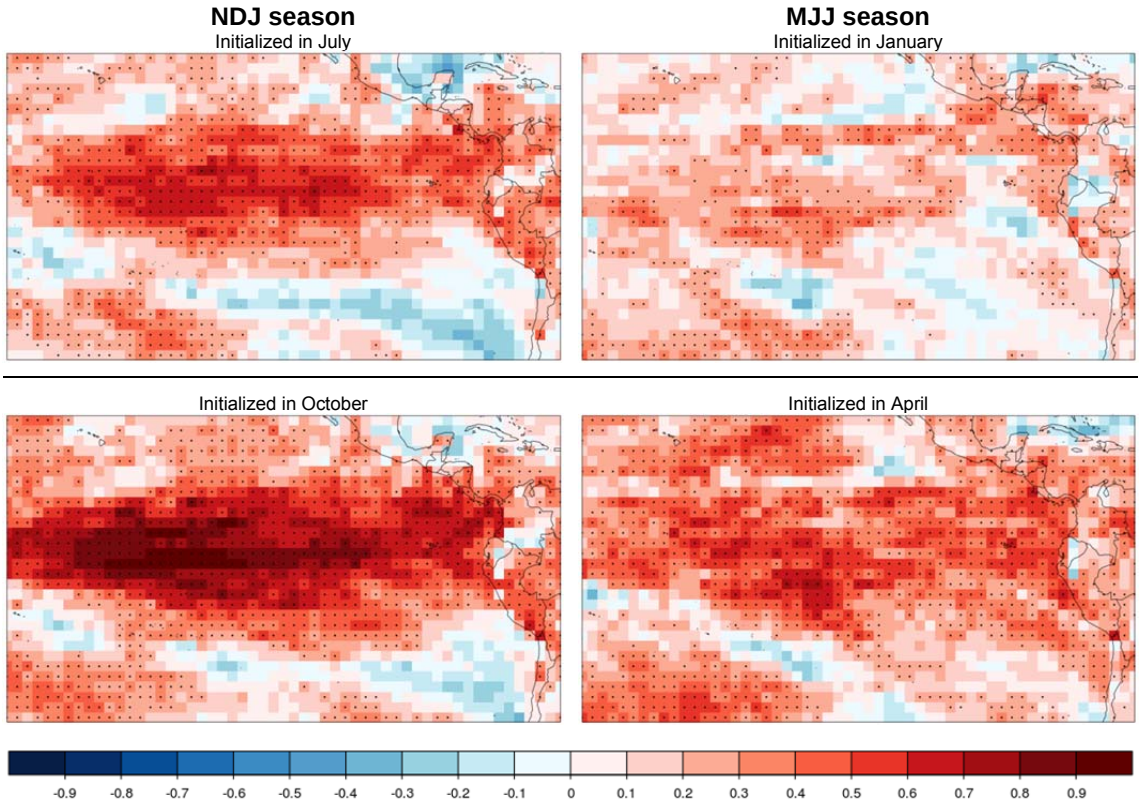

Figure 4: Fair Ranked Probability Skill Score for seasonal temperature forecasts in the ENSO region. The plots show the area between 170 °W to 70 °W and 30 °S to 20 °N. The forecasts on the left are issued for the NDJ season and to the right forecasts for the MJJ season are shown. The lead time decreases from top to bottom so that forecasts with lead times 5 to 7 months are shown at the top and lead times 2 to 4 months at the bottom. Positive values indicate that the forecast outperforms a constant climatological forecast and significantly (at the 5% level) positive RPSS are stippled.

This difference in predictability for different seasons can be explained by the Spring Predictability Barrier (SPB). ENSO is a strong signal of climate variability but during the spring months even making a forecast for summer is difficult. In their study, Torrence and Webster (1998) show that the SPB is seen as a rapid decline of persistence for the months March, April and May. They find that during this time of transition between ENSO periods the signal to noise ratio is lowest and the susceptibility to perturbations is highest. Furthermore, they state that the annual cycle in the eastern Pacific shows a warming during winter and spring, which makes it even harder to identify a potential El Niño. Other studies similarly suggest that the SPB is an intrinsic property of ENSO and explain the SPB phenomena by weak ocean-atmosphere coupling (Webster, 1995) and the low SST anomaly signal during spring (Xue et al., 1994). However there is still an ongoing debate on this topic and there are numer-

ous studies indicating that the SPB can be reduced by addressing initial errors and model errors (Chen et al., 1995; Mu et al., 2007; Duan et al., 2013). In a more recent study, Zheng et al. (2010) use an Ensemble Prediction System to show that both the initial error degrading forecast skill in spring and the small predictable signal are main factors contributing to the SPB.

In our skill analysis the reduced predictability due to the SPB can also be found in forecast discrimination. The two plots of the top panel in Figure 5 show the Generalized Discrimination Score for the same region and target seasons. Forecasts have less discrimination in the MJJ season, which is true for lead times of 5 to 7 months (plots at the top) and shorter lead times (see Figure A-1 in the appendix). This indicates that during boreal spring, forecasts have less ability to distinguish outcomes for which observations differ. For comparison, we show the ROC area score in the middle and bottom panels of Figure 5 for the same target seasons and lead time but only for categories 1 and 3, which correspond to the coolest and the warmest one-third in this case. Note that the color scales for the two measures of discrimination are different since the ROC area score is a skill score, calculated by comparing the discrimination of the forecasts to a reference.

The ENSO signal stands out more for the ROC area score than for the Generalized Discrimination Score because the ROC area score divides the forecasts into categories and we look at the lower and upper terciles only. The middle category, which has lower predictability and more background noise from sampling uncertainty, is not shown. The NDJ season has higher skill than MJJ for all lead times and terciles (see also Figure A-2 in the appendix). For NDJ there is slightly higher discrimination for the warmest one-third (upper tercile). This would indicate higher predictability of the warming phase (El Niño) compared to La Niña. But the difference is not very obvious because the area of highest skill is shifted eastward to the South American coast. On the other hand the MJJ forecasts clearly show a better skill for the coolest one-third, especially for forecasts with longer lead times. This indicates that it is harder to distinguish a warming phase from normal conditions (or colder than normal) for spring forecasts than predicting a cooler than normal spring. This supports the hypothesis by Torrence and Webster (1998) that the warming of the eastern Pacific during spring makes it even more difficult to identify a warmer than normal signal.

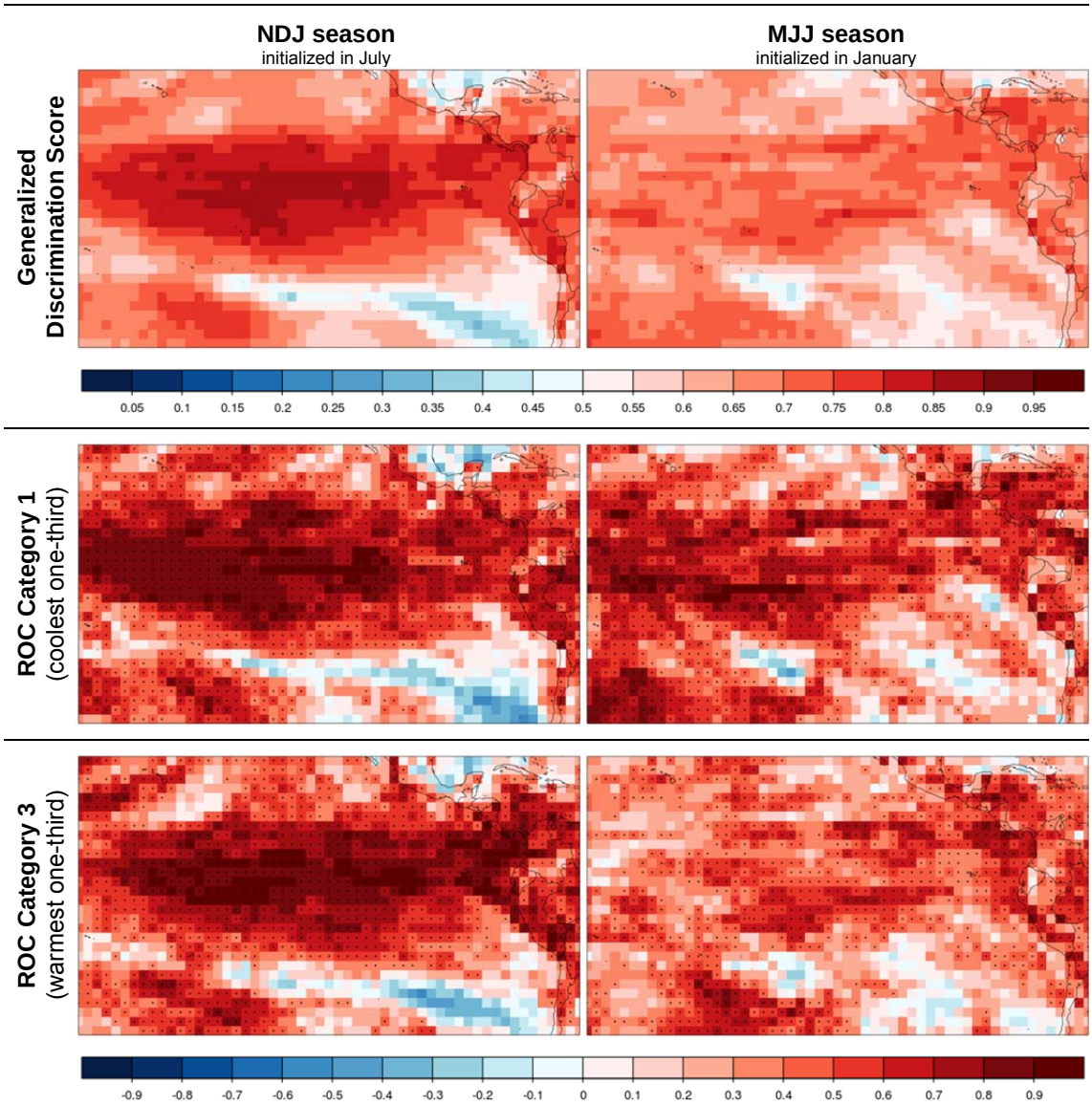

Figure 5: Same as Figure 4 but for measures of forecast discrimination Shown in the two plots of the top panel is the Generalized Discrimination Score: Blueish colors indicate a score below 0.5, which means that the forecast has no discrimination and performs worse than guessing in telling different cases apart. Reddish colors indicate that the forecast has discrimination. Middle and bottom panel: ROC area score showing skill of predicting the coolest one-third (two plots in the middle panel) and warmest one-third (two plots in the bottom panel). Positive values shown in red indicate that the forecast outperforms climatology. Forecasts significantly (at the 5% level) better than guessing the category are indicated by stippling of the respective grid cell. Forecasts for the NDJ season are shown to the left and for the MJJ season to the right. All plots show forecasts with lead times 5 to 7 months, see Figure A-1 and Figure A-2 in the appendix for lead times 2 to 4 months.

In the annual cycle the SPB can be clearly seen. Figure 6 shows the monthly Fair RPSS for a grid point in the Niño3.4 region. The plot at the top shows all lead times by target month (on the x axis). The forecasts have skill over climatology for all lead times and target months since no values are zero or smaller. A drop in accuracy can be seen for all lead times in March and forecast performance is worst in April and May. Comparison with other grid points in the area confirms that lowest Fair RPSS is found in March, April and May (not shown). Furthermore it is obvious from Figure 6 that best accuracy values are obtained for the winter months November, December, January and February. In the bottom plot skill is plotted against lead time. Lines of longer lead times mostly lie below lines of

shorter lead times, indicating that in general skill is lower for longer lead times. Exceptions are forecasts for March, April and May where skill oscillates roughly around the same level for all lead times. For forecasts initialized in October, November and December the SPB is seen as a steep decrease of skill for target month April (triangles in Figure 6 mark target month April). We can even identify examples for increased skill after the SPB: The forecasts initialized in March and April have higher skill for longer lead times (target months June and July) than for lead times 1 and 2 months. Also, forecasts for March are better for longer lead time or earlier initialization (which is true for lead times 4 to 6 months compared with 1 to 3 months; see e.g. top panel in Figure 6, highest skill in March for initialization months October and December, which corresponds to lead times 4 and 6 months respectively). However we have to keep in mind that this is just an example for one grid point.

We take from this that forecast skill in this region depends on whether the forecast are issued before or during spring (which would then mean that forecasts are less skillful due to the SPB). Time of forecast initialization and lead time together are essential to explain forecast skill.

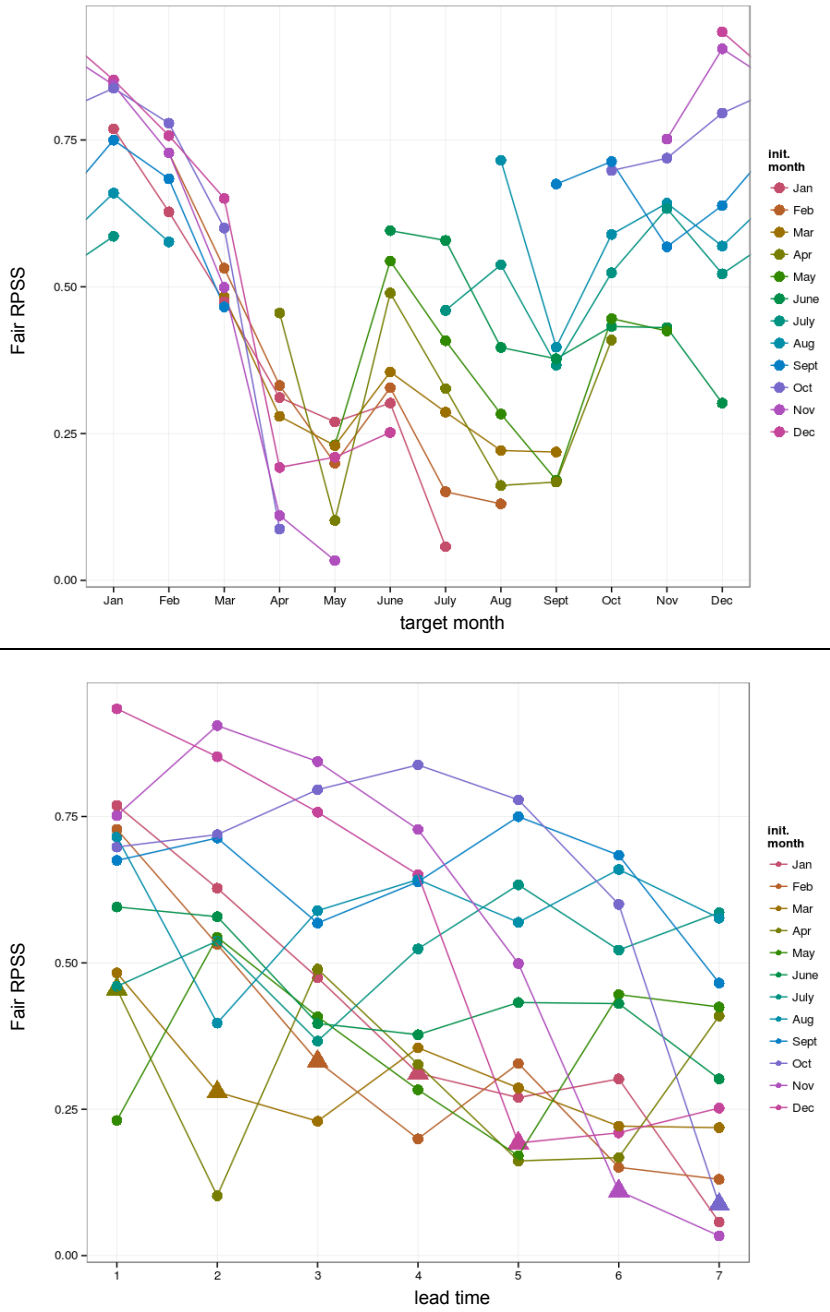

Figure 6: Monthly Fair RPSS for temperature forecasts at a grid point (152 °W, 4 °S) in the Niño3.4 region. In the panel at the top all lead times are shown by target month. Forecasts initialized in the same month (init. month) are connected with lines. In the bottom plot skill is plotted against lead time, again forecasts with the same initialization month are connected with lines. April as the first month after the Spring Predictability Barrier is marked with a triangle in the lower panel. The scale on the y axis is the same for both plots.

3.2 Fall precipitation forecast for Indonesia

Rainfall in the Indo-Pacific region occurs throughout the year with peak rainfall from December to February, which coincides with the Northeast monsoon (also known as the Northwest monsoon in the Austral-Indonesian region; Hendon, 2003). Potential predictability of precipitation during this season, however, was found to be low (Hastenrath, 1987; Haylock and McBride, 2001; Hendon, 2003), whereas the monsoon onset and rainfall during the transition season (September to November) is potentially predictable (Nicholls, 1981; Hastenrath, 1987; Haylock and McBride, 2001).

It seems that the monsoon onset date is potentially predictable due to the strong relationship in that season with ENSO (Robertson et al, 2008). Warm ENSO events induce delay of the onset of the southern hemisphere summer monsoon, as was shown for example by Moron et al. (2008). Through the timing of the monsoon onset, predictability of rainfall in the Indo-Pacific region is related to the Southern Oscillation Index and thus influenced by the ENSO cycle (Hastenrath, 1987; Haylock and McBride, 2001; Moron et al., 2008). Previous studies have further shown that the ENSO-precipitation connectivity is highly variable, spatially and seasonally. It is strongest during the northern hemisphere fall (SON) when an ENSO event is in its growth phase and it gets weaker during winter (DJF) when ENSO matures (Haylock and McBride, 2001; Tangang and Juneng, 2004; Juneng and Tangang, 2005). Tangang and Juneng (2004) specify that ENSO-related coherence covers the whole region of islands in the Indo-Pacific during SON but it migrates northward during DJF, which in fact induces a strengthening of the relationship with rainfall in Malaysia while the relationship weakens for the large part of the region in the Southern hemisphere. In the following the focus will be put on seasonal forecast skill of fall and winter precipitation in the Indo-Pacific region.

Above features of predictability of rainfall in Indonesia and the Indo-Pacific are confirmed in the skill assessment presented here and illustrated for fall (SON) and winter (DJF) forecasts in Figure 7. The graphics show forecast skill measured by the Fair RPSS for all seasonal forecasts between August and March in the Indo-Pacific region. Indeed forecast performance is best in the SON season and seems to decrease thereafter. For the seasons starting one month ahead of and after SON (August-October and October-December) high forecast accuracy is found as well but for a smaller region (see Figure A-3 in the Appendix). During NDJ forecast skill is visibly decreasing and there is no skill of precipitation forecasts for the Indo-Pacific islands for DJF and JFM (see Figure A-3 for DJF). The only exception is found for the Philippines and partly Malaysia, which agrees with the theory of a northward shift of predictability during DJF (Tangang and Juneng, 2004).

The onset of the Asian summer monsoon does not occur simultaneously across the entire Indo-Pacific region but takes place from northwest to southeastern direction during late August to mid-December (Tanaka, 1994; Moron et al., 2008). The observed large-scale mean onset date of the rainy season is sometime in October. Moron et al., 2008 found that the predictability of the onset across the region is highly correlated with the July SST in the ENSO region, which indicates skillful predictions over several months lead time. Looking at the skill of forecasts of monthly precipitation highest correlation for the entire region can be found in October (see Figure 8). Again a northward shift of skill can be seen starting in November and proceeding first towards Brunei, Malaysia, Vietnam, and, later on, mainly to the Philippines.

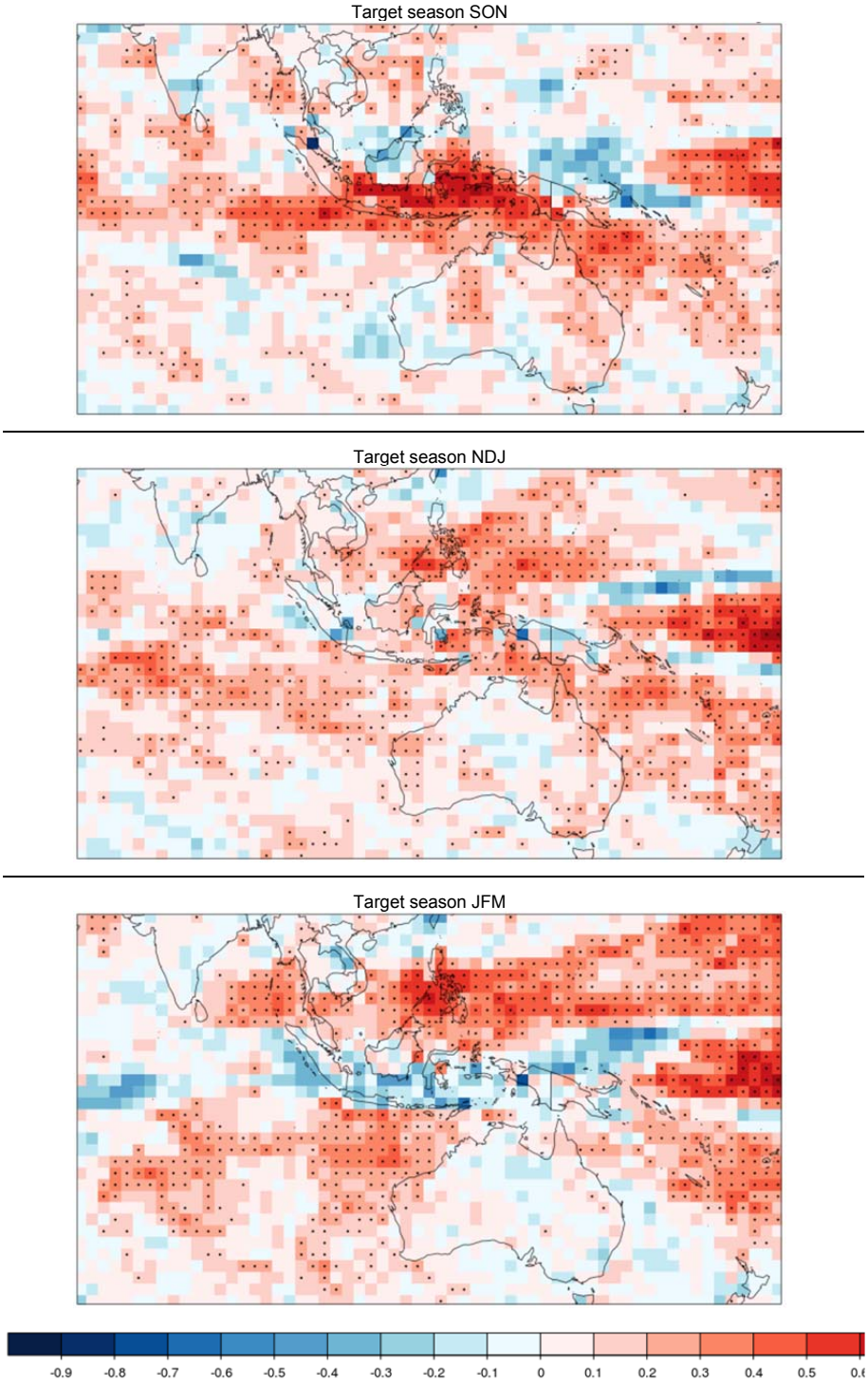

Figure 7: Fair RPSS for seasonal precipitation forecasts in the Indo-Pacific including Australia. Forecasts for lead months 2-4 initialized in August, October and December are shown.

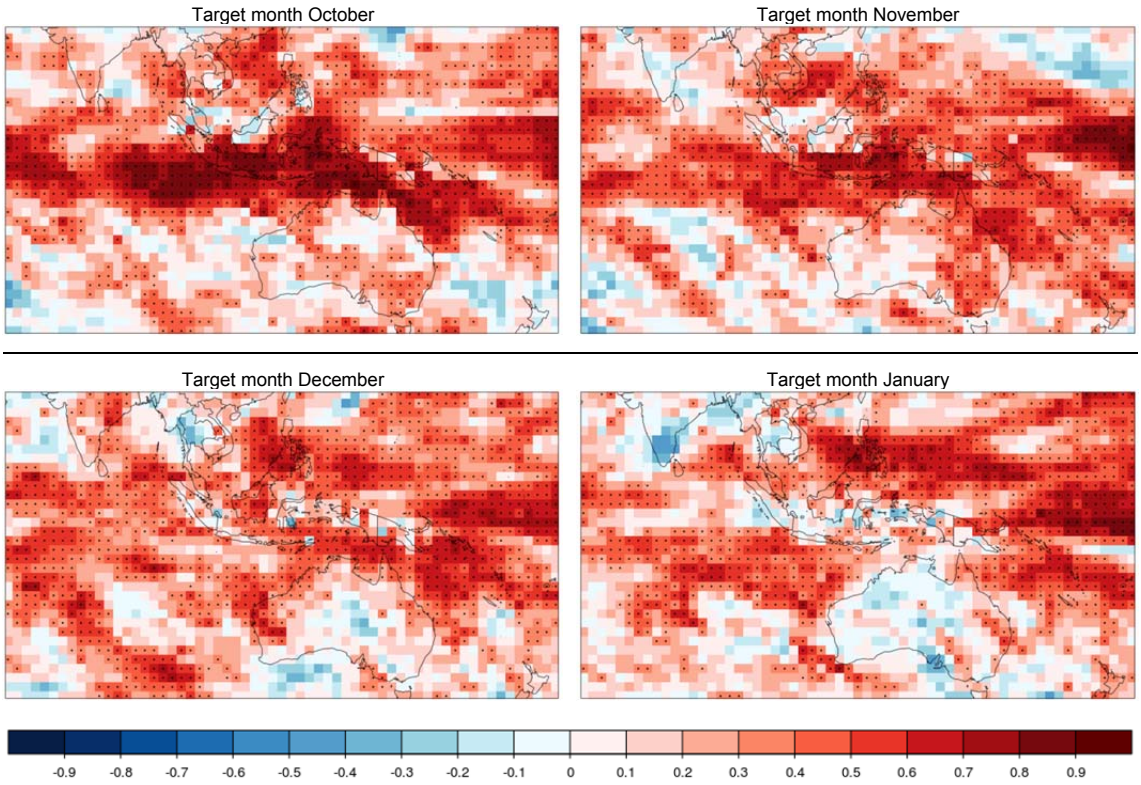

Figure 8: Correlation coefficient for monthly precipitation forecasts in the Indo-Pacific region. Forecasts have lead time 2 months, which means that the forecasts were initialized between September (for target month October) and December (for target month January).

Figure 7 and Figure 8 show forecasts with a lead time of 2 months. When comparing precipitation forecasts with longer lead times it seems that for this region of the world the dependence on target month is much stronger than on lead time. This is shown exemplary for seasonal forecasts measured by the Fair RPSS in Figure 9 and in Figure 10. There is skillful prediction of fall (SON) precipitation over long lead times for most of Malaysia and Indonesia (Figure 9). On the other hand, the Fair RPSS for all lead times shows roughly a constant level of lower forecast accuracy for predictions of winter (DJF) precipitation in the same region. The same can be observed for forecasts of monthly averages and other verification metrics (not shown).

Dependence on target month is demonstrated for a representative grid point in the Indo-Pacific Ocean located south of the Indonesian province South Sulawesi in Figure 10. Seasonal forecasts covering the summer months comply a little more with the intuitive assumption of decreasing forecast skill with lead time (see e.g. JJA and JAS in the bottom plot). The top plot in Figure 10 shows that the fall months, as in ASO, SON and OND, are easier to predict accurately compared to seasons covering winter months as in DJF and JFM. The lines of equal lead times are close together for all target seasons and none of the lead times has always the lowest or always the highest skill. A barrier to forecast skill based on the initialization time as would be given for example by July SST was not found. Skill increases continuously after JJA to its peak in SON (Figure 10, top).

In Figure 11, different verification metrics are shown for the September precipitation forecast in the Indo-Pacific. In the first pair of panels we find strong correlation between the forecasts and the ob-

servational dataset in central and eastern Australia. When examining forecast accuracy, however, the forecasts do not seem to perform as well as one would think based on the association (compare also with the Fair RPSS in Figure 7 and Figure A-3). The plots illustrate that the strong correlation is not supported by either of the accuracy measures. For both lead times shown in Figure 11 the Fair RPSS and its extension for continuous forecasts, the Fair CRPSS, indicate low skill of the forecasts mainly in central Australia but also the eastern part. The same applies to Papua New Guinea. The Fair CRPSS is a more strict skill metric than the Fair RPSS since it assumes an infinite number of categories (instead of the 3 chosen here for the Fair RPSS) and does not account for biases. However, note that the scale of the skill scores goes from minus infinity to one. This means that the skill is lower when measured by the Fair CRPSS but the dark blue color seen in Figure 11 in the Pacific north of Papua New Guinea is only an artifact of the color scale varying between -1 and 1.

The low forecast accuracy we just identified for parts of Australia and Papua New Guinea is partly explicable with the other skill metrics (Discrimination and ROC area score (not shown) indicate low forecast discrimination for Australia) or might be related to single extreme values. The plots at the bottom of Figure 11 show a measure of forecast reliability, the Fair Spread to Error Ratio. Most of the area around Indonesia and Papua New Guinea appears in blue colors, which means that the ratio is smaller than 1. It seems that one of the problems with forecast performance in this area is overconfidence, which is also found for temperature forecasts (not shown).

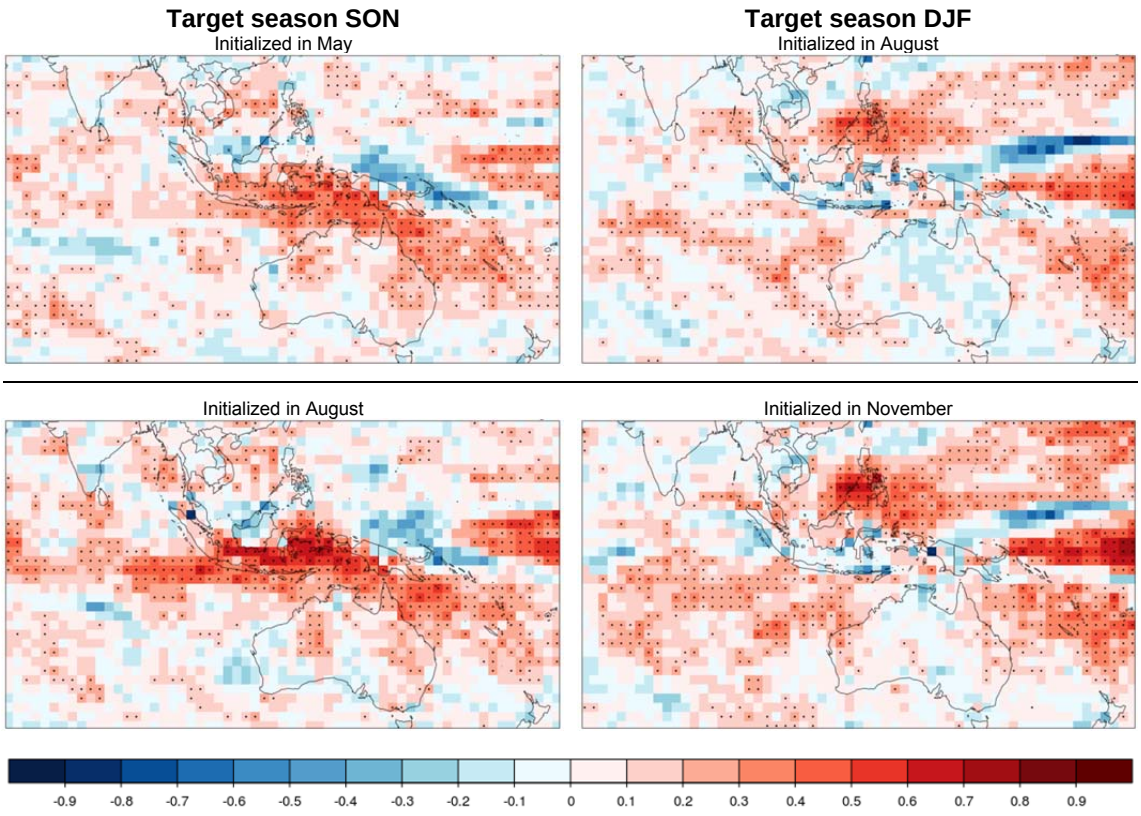

Figure 9: Forecasts skill in the Indo-Pacific region measured by the Fair RPSS. The forecasts to the left are for target season SON and to the right for target season DJF. Shown are lead times 2 (plots at the bottom) and 5 months (plots at the top) ahead the target season.

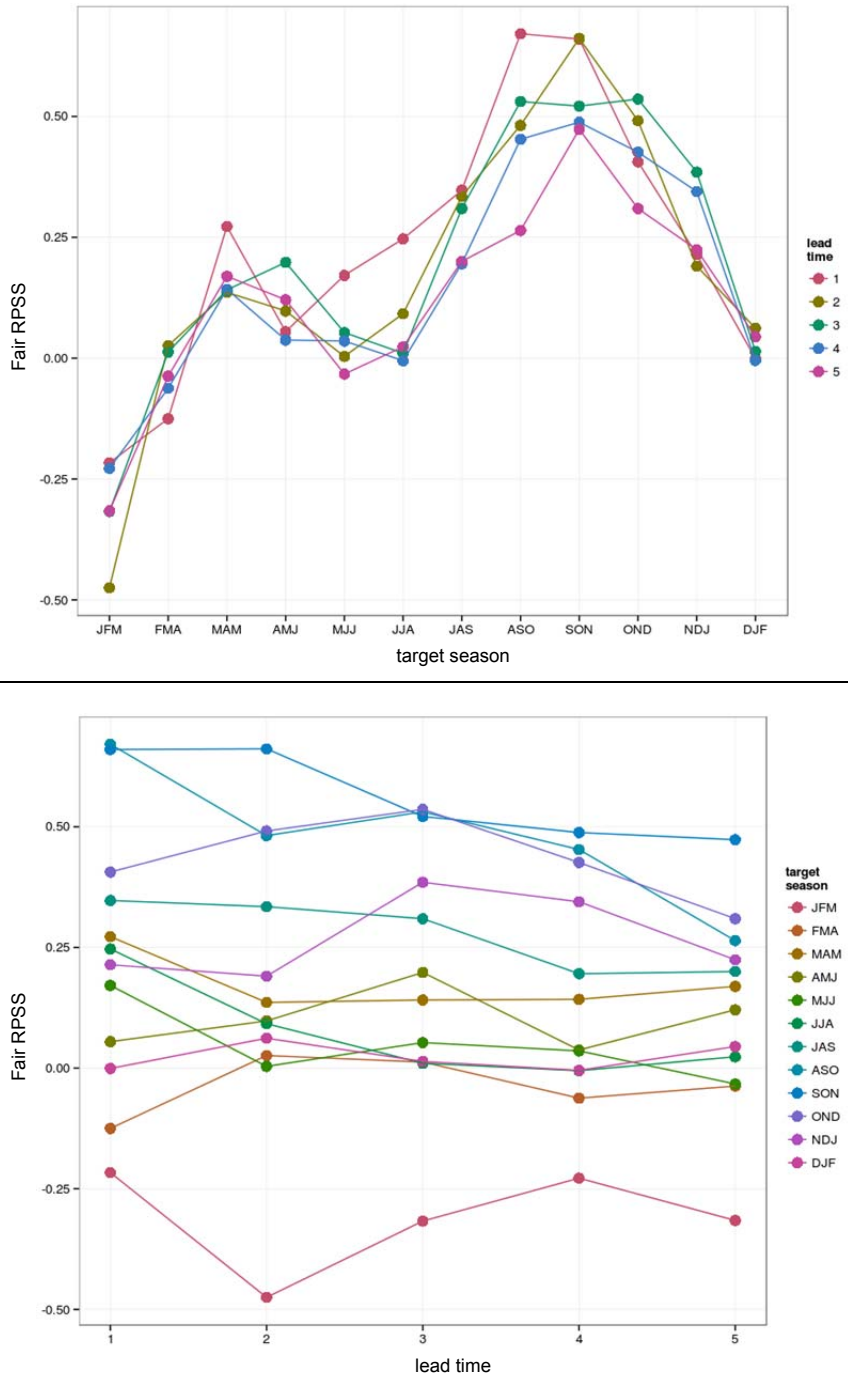

Figure 10: Seasonal Fair RPSS for precipitation forecasts at a grid point (120 °E, 6 °S) in the Indo-Pacific. The plot at the top shows all lead times by target season and the plot at the bottom all target seasons by lead time. The scales on the y axis are the same for both plots.

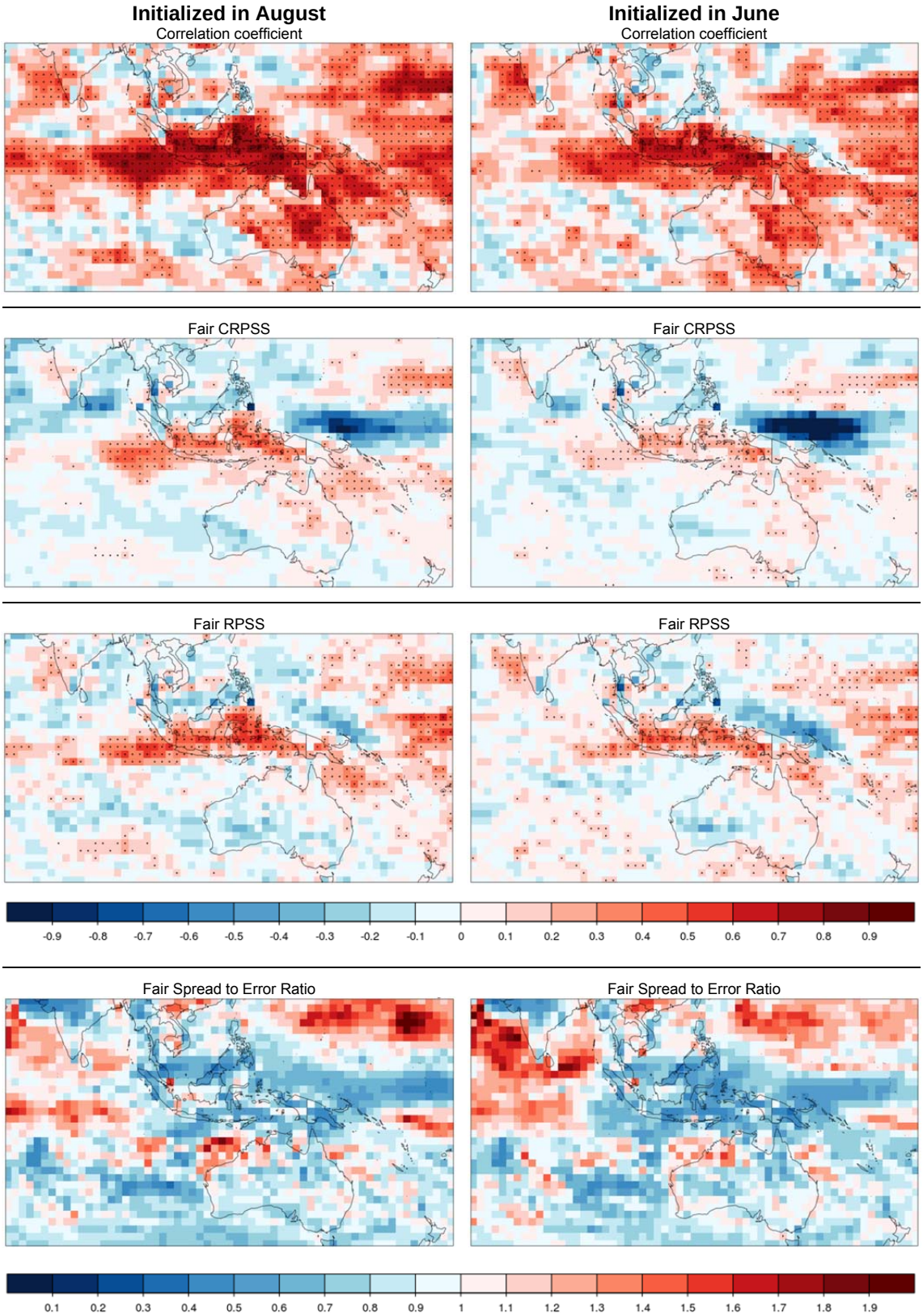

Figure 11: Various verification metrics for monthly precipitation forecasts in the Indo-Pacific region. The forecasts are initialized in August (left) and June (right) for target month September (lead times 2 and 4 months). Please note that the Fair Spread to Error Ratio has a different color scale from the rest. Spread to Error Ratio larger than one indicates overdispersion, smaller than one indicates overconfidence of the forecasts.

3.3 European winter forecast

Similar to ENSO, which is the most important driver of climate variability in the tropics, variability of the winter surface climate in some areas of the Northern extratropics is governed by the state of the North Atlantic Oscillation (NAO). Walker and Bliss (1932) defined an index to characterize the state of the NAO based on the sea level pressure difference between stations in Iceland and the Azores. These locations were chosen due to a high anticorrelation, related to low pressure systems located at Iceland and high pressure systems over the Azores at the same time (the typical NAO dipole pattern). Essentially, the NAO index describes the meridional pressure gradient over the northern Atlantic and consequently the NAO is a proxy for the strength of the westerly winds across the North Atlantic (Serreze and Barry, 2005).

The positive mode of NAO is characterized by a strong meridional pressure gradient. This implies stronger than normal westerly winds across the Atlantic and a northward shifted jet. During positive NAO conditions northern Europe receives warm and moist air and severe storms and heavy rainfall are more likely to occur, while in turn the occurrence of extreme low temperatures is reduced (shown e.g. by Smith et al., 2016). In the negative NAO mode the meridional pressure gradient is weak and this leads to suppressed westerly flow. As a consequence cold and dry continental air is frequently advected to large areas of northern Europe. The storm tracks move more to the south and bring moist and warm air masses to the Mediterranean. In extremely negative NAO conditions, the pressure gradient can be reversed, leading to an easterly wind across the Atlantic (Smith et al., 2016).

As the winter climate in Europe is driven by the state of the North Atlantic Oscillation (NAO), we hypothesize that predictability of the NAO would lead to skillful winter forecasts in Europe. Evidence that there is predictability of NAO and the related climate impacts on seasonal time scales was found in recent studies using the UK Met Office seasonal forecasting system (e.g. Scaife et al., 2014; Smith et al., 2016). Although sources of NAO predictability are not always well represented in forecasting systems, these studies have shown that present day seasonal forecast models may achieve useful levels of forecast skill. Other studies show differing results though. Müller et al. (2005b) find skill in the NAO forecasts of ECMWF System 2 as well, but attribute it to the small number of years being considered. Shi et al. (2015) find indications for low frequency variations in predictability of the NAO investigating 40 years of hindcasts. Later studies further confirm these results using data sets covering the whole 20th century (Weisheimer et al. 2016). Whether or not this is the case for ECMWF System 4 is yet unknown and we will therefore investigate levels of predictability of winter precipitation and temperature forecasts in the following.

In Figure 12 we show the correlation coefficient for seasonal forecasts for the four standard seasons. Highest correlation is found in central and northern Europe in spring (MAM) for short lead times. In summer (JJA), there is a drop of skill in central Europe while correlations are highest in summer for Eastern Europe and the Mediterranean. Skill recovers during fall (SON), in particular for Great Britain and Northern Ireland. Correlations for winter (DJF) are generally lower than for the other seasons. During winter and spring, we find pronounced variability in correlation with lead time, but forecasts at shorter lead times do not necessarily exhibit higher correlations, as can be seen for the monthly winter forecasts (Figure 13). Forecast correlation is found to be particularly low for December and January with still considerable variability for forecasts with varying lead time. The variability in forecast skill and also the skill not being significant in many areas could indicate a high noise level due to sampling error. This indicates that either the skill of NAO forecasts in System 4 is limited or that the

teleconnections are misrepresented leading to an erroneous NAO-related signal over the European continent and thus limiting forecast skill in winter.

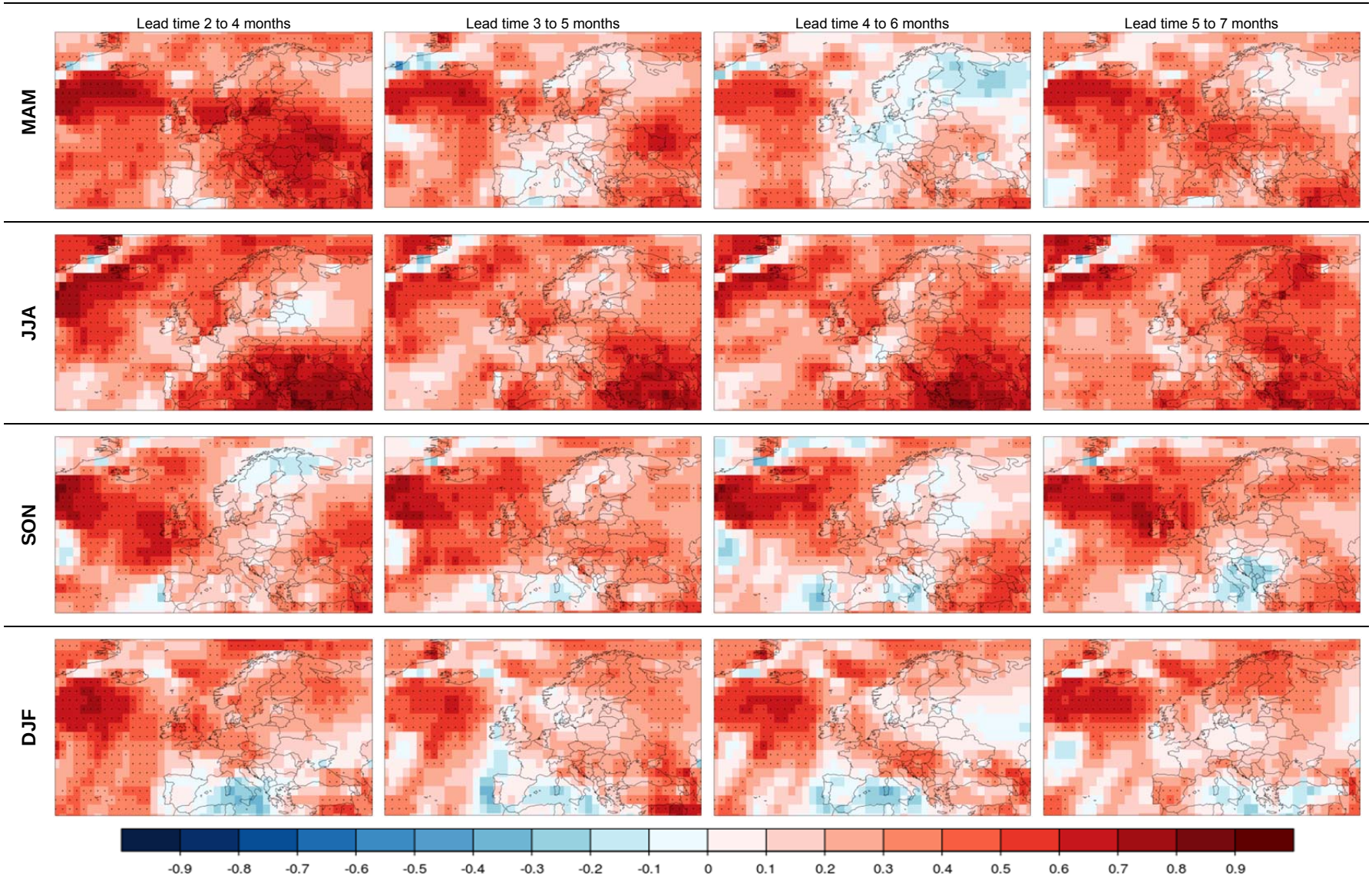

Figure 12: Seasonal temperature forecasts for North Atlantic and Europe verified using the Correlation coefficient. Shown are lead times [2,3,4] to [5,6,7] months (from left to right).

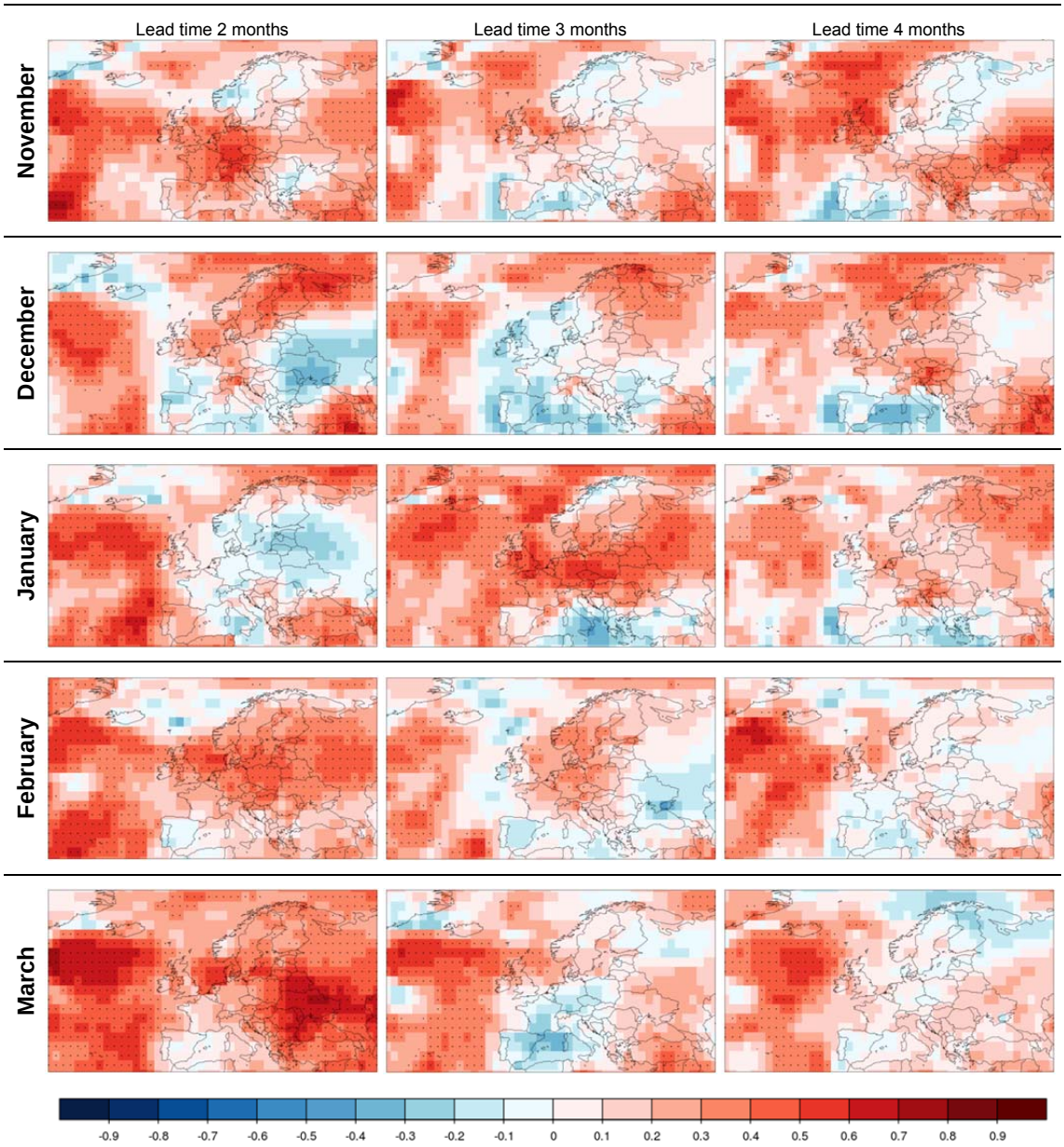

Figure 13: Correlation coefficient for monthly forecasts for November to March and lead times 2 to 4 months.

In the following, we illustrate how such a forecast quality assessment could be used to address user-relevant questions. For this we focus on snowfall forecasts for the Christmas holidays that are highly relevant for the tourism sector. Snowfall forecasts require both temperature and precipitation forecasts with sufficient forecast quality.

In Figure 14, we show the Generalized Discrimination Score for December temperature and precipitation forecasts initialized in September (lead time 4 months). While discrimination for temperature is positive, we find little discrimination for precipitation. Discrimination in precipitation does not improve for any of the shorter or longer lead times (not shown), which indicates that forecast performance is not dependent on lead time but determined by the target month and variable. The Discrimination Score for precipitation is between 0.65 and 0.7 at best, which is barely better than guessing out-

comes. The lack of skill also explains why seasonal precipitation forecasts for European winters are issued with reluctance on an operational basis to the public.

Discrimination for temperature forecasts is generally higher (left panel in Figure 14) and positive for most of the region with maximum values in a region close to the Adriatic Sea covering Slovenia and Croatia. Yet again discrimination is not strong, with values below 0.7. For other lead times, discrimination is even lower. From this we conclude that we do not expect forecasts for the Christmas holidays issued weeks to months ahead to be skillful. This applies to temperature and precipitation forecasts and consequently to snowfall determined by these variables.

Comparing the Generalized Discrimination Score (Figure 14) with the correlation shown in Figure 13, we also note that the two verification metrics behave similarly.

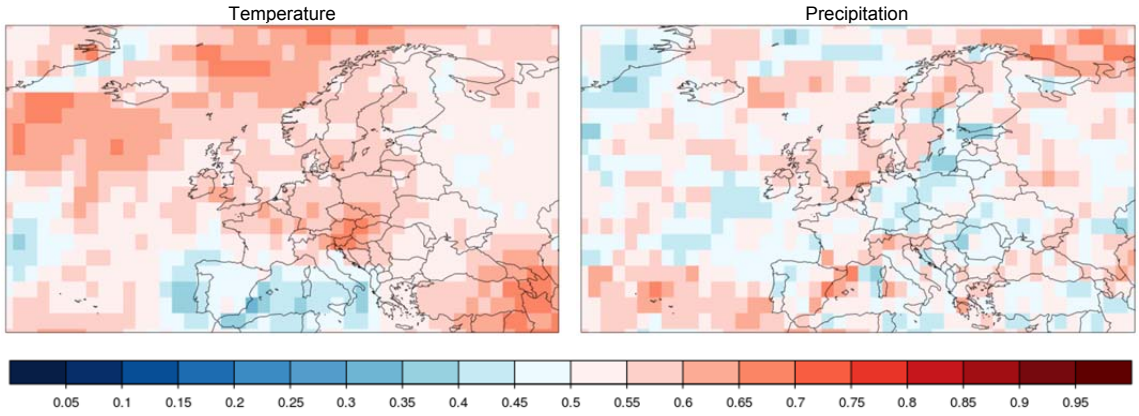

Figure 14: Discrimination of temperature (shown on the left side) and precipitation (right side) forecasts measured by the Generalized Discrimination score. Shown are forecasts for December averages initialized in September (lead time is 4 months).

Forecast skill for December temperature forecasts in central Europe can be explained by the results found for forecast reliability shown in Figure 15. The high values found for the Fair Spread to Error Ratio in Central Europe indicate overdispersion in the temperature forecasts to be one of the main issues. The same pattern is found for all lead times. The tendency to overdisperse might relate to one of the results from Scaife et al. (2014) for the seasonal forecasts of the NAO using the system of the UK Met Office (Global Seasonal forecast System 5). The authors find that the predictable signal in the NAO forecast is weaker than in the observations. As a consequence, they hypothesize that increased ensemble size would increase seasonal forecast skill due to the increase in signal-to-noise ratio of the forecast signal. Whether this also applies to ECMWF System 4 forecasts would require further analysis, which is beyond the scope of this technical report. It must also be noted, that the pattern found here does not relate directly to the temperature pattern of the NAO.

In contrast to temperature, the precipitation forecasts seem neither overdispersive nor overconfident. Building on the findings up to this point it seems that there is low potential predictability for precipitation, a conclusion, which is supported by all skill metrics. Also based on the skill assessment of temperature and precipitation forecasts, we do not expect that there is potential for skill improvement of monthly or seasonal snow forecasts for Europe.

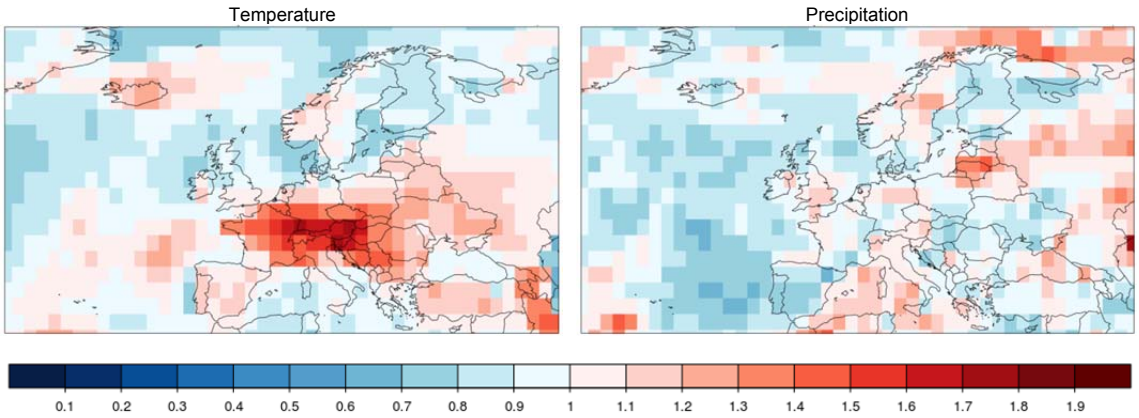

Figure 15: Fair Spread to Error Ratio for temperature (shown on the left side) and for precipitation (right side) forecasts. Shown are forecasts for December initialized in September (lead time is 4 months).

The seasonal outlook for temperatures of the whole winter looks similar to the December discrimination and reliability results shown in Figure 14 and Figure 15 (not shown). Again the forecast is better in discriminating among observations than guessing the outcome but there is indication that overdispersion is one of the main issues why the forecast signal is not very strong. Comparing these findings with one of the operational products of the ECMWF System 4 model (see Figure 16) we can understand why there is no clear trend in seasonal forecasts for the 2016 European winter temperatures.

In Figure 16 we show the operational forecasts for 2m temperature initialized on 1st of October 2016 for the following NDJ season. In the raw forecasts (left panels) there is a small tendency towards the upper tercile for most parts of Europe and most pronounced in the south. In the raw configuration over 42% of the ensemble members predict a higher than normal temperature compared to climatology (from 1981-2015), which is indicated by the light red color. Consequently less than one third of the ensemble members predict a colder than normal NDJ season (top left panel). Recalibration scales the ensemble mean forecasts and the ensemble spreads in order to make forecasts reliable (see e.g. Weigel et al., 2009 for the climate-conserving recalibration). The recalibrated forecasts do not show a tendency towards either of the terciles (except for a signal against the lower tercile around northern Italy and the Adriatic Sea that remains from the raw forecast). Recalibration reduced the tendency towards the warmer tercile and against the cooler tercile by scaling of the ensemble spread. As a consequence the recalibrated forecast has no clear trend.

For the NDJ precipitation in Europe the operational forecast shows no trend for the current year (not shown). This is explained by the low potential predictability in winter we find in this skill assessment.

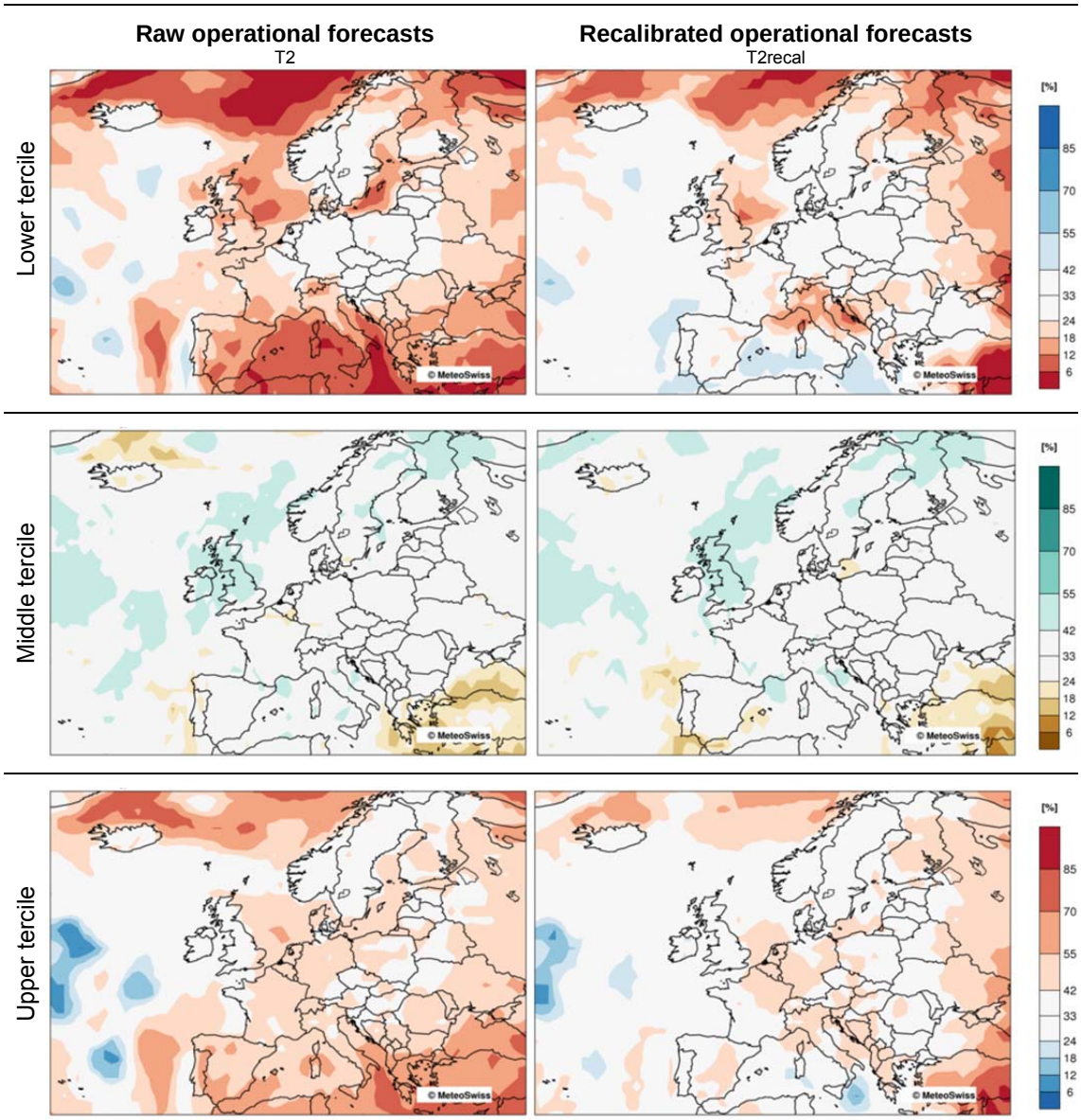

Figure 16: Operational forecasts from the ECMWF Seasonal forecasting system produced by MeteoSwiss. Forecasts for 2m temperature are produced from 51 ensemble members initialized on October 1st, 2016 for target season NDJ. Shown is the probability of 2m temperature being in the lower tercile (top), middle tercile (middle) or upper tercile (bottom). The model climatology is from 1981 to 2015. The panels to the left show the raw operational forecasts (T2) and the panels to the right the recalibrated forecasts (T2recal).

3.4 Forecast verification artefacts

Mainly for precipitation forecasts, one can find unexpected patterns that are difficult to interpret from a physical perspective. For example, grid cells with significant and high forecast skill are found next to grid cells without any skill. This results in a chess-board-like patterns as can be seen for precipitation forecasts in the central Andes in Figure 17. A possible explanation for this pattern in the central Andes relates to the way these mountains are represented in the model. The steep gradients in this area are likely to dominate the rainfall regime and potentially even induce wave-like rainfall artefacts that negatively affect skill in certain grid cells (Tim Stockdale, personal communication). Not only model but also reanalysis topography was found to differ from reality, leading to displacement of rainfall patterns compared to observations (Katrin Sedlmeier, personal communication). Empirical statistical or dynamical downscaling of the seasonal forecasts could provide further insight into this question. Also, remember that we mentioned in chapter 2.1 that the reliability of the reanalysis we use as the verifying observations depends on the quality and availability of observational data.

In the figures that we showed for precipitation in Indonesia we can identify the same kind of artefacts (see e.g. Figure 7 and Figure 9). In this case the explanation could be differences in the land-sea mask of the model and the reanalysis.

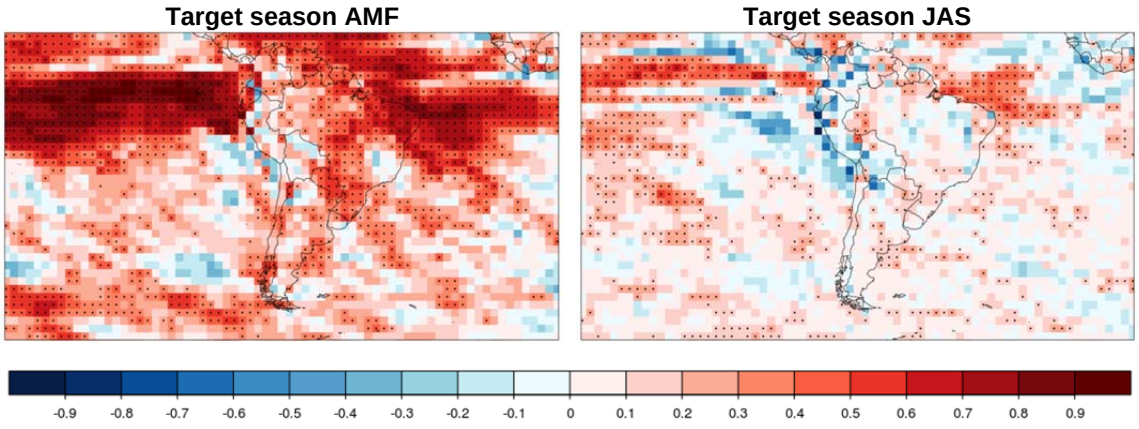

Figure 17: Chess-like patterns seen for seasonal precipitation forecasts in South America. The plot to the left shows the correlation coefficient for the AMJ season initialized in March .The plot to the right shows the Fair RPSS for the JAS season initialized in June. Note that both skill metrics shown have the same color scale. However, the meaning is different.

In Africa, in the area of the Saharan desert, similar patterns can be found, (Figure 18). In this region of the world, where rain events are very rare, the episodic nature of precipitation provides a possible explanation. Also the short time period of only 34 years of data used for the forecast quality assessment might be insufficient to capture the rare and episodic occurrence of rainfall and multi-decadal variability.

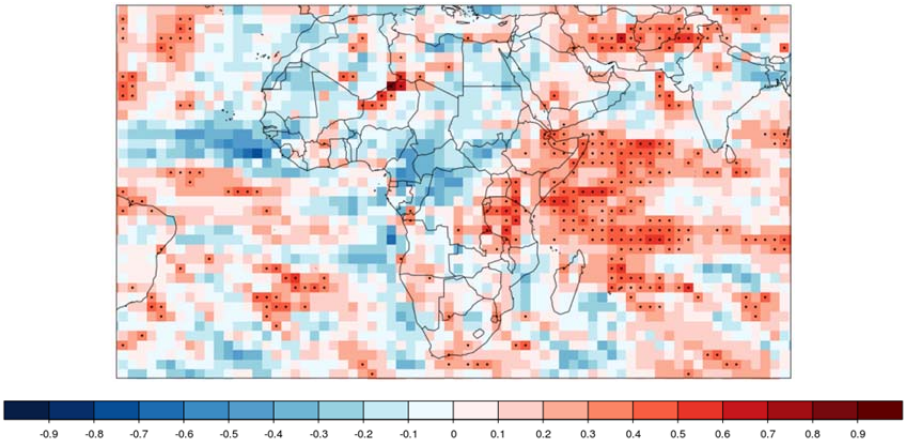

Figure 18: Correlation coefficient for monthly precipitation forecasts in Africa displaying chess-board-like pattern in the Saharan Desert and the Sahel. Forecast is initialized in June for target month November.

4 Visualization of forecast skill

To engage with users and to explore the space-time variability of seasonal forecasts skill interactively, a web application was developed using the R programming language with the Shiny for R package. The app uses the results of the forecast verification to create skill maps on-the-fly. Online access is granted with the help of shinyapps.io, a hosting service by RStudio. In this chapter first the tools are described and then the app and its functionalities are presented.

4.1 Shiny for R

Shiny is an R package developed by RStudio that serves as a web application framework. It allows R users to build interactive applications straight from R and does not require knowledge about any web-related programming language. Nevertheless the appearance of the Shiny app can also be modified flexibly using HTML, CSS and JavaScript commands. In fact a part of the Shiny functions can be used interchangeably with common HTML tags and Shiny functions understand HTML attributes passed to them. The Shiny package can be installed directly in R through the command line[1]. The full documentation of the package including tutorials and live examples is available online: http://shiny.rstudio.com.

A Shiny application is a web page connected to a computer running a live R session. Users can set inputs and perform manipulations on the web page, which will cause R code to be run and the content of the web page to be updated. The basic structure of a Shiny application has two components:

- User-interface (UI) script **ui.R**: controls the layout and appearance of the app and is mainly built with HTML code that is nested in R functions.
- Server script **server.R**: contains instructions to build and rebuild the R objects displayed in the UI (including general R code and interactivity directions).

Each Shiny application needs its own directory where the ui.R and server.R files are saved together with other files and data accessed by the app. When running the application the ui.R and server.R are combined into a functioning app that can access all the other sources in the directory. The app can be run as an internal web page in RStudio or in a web-browser connecting to a Shiny server (either locally or over the internet). Interactivity is created by input and output functions that save and display information using the corresponding variables. The information is carried between the ui.R

1 Run install.packages(„shiny“)

and server.R using these reactive input and output values. Outputs are built by reactive calls that re-execute automatically when users modify inputs.

4.2 Deployment of Shiny applications

Shiny applications can be run within RStudio and the source code can be run on any other machine that has R installed. Sometimes, however, it is desired to make the application available to a broader group of users that do not have access to an R and RStudio installation locally. Also, one may wish to make the interactive web application publicly available without sharing the R code used to generate the app. The easiest solution is to share the application on the web. Shiny applications are designed to work on the web either as standalone websites or embedded in a parent website. To deploy the app on the web one can set up a server or host the app in a cloud-based service using the https://www.shinyapps.io platform.

Shinyapps.io is a hosting service that runs in the cloud on shared servers operated by RStudio. This implies that the application is outside the firewall of the creator. Furthermore the data that is used by the application is uploaded and available to the cloud (but not accessible to the user other than through the output displayed in the app). Access to the application is encrypted with SSL (Secure Sockets Layer), a cryptographic protocol ensuring secure communication over a computer network (Dierks and Rescorla, 2008). The user guide and specifications for shinyapps.io can be found on the web: http://docs.rstudio.com/shinyapps.io.

To start with shinyapps.io and to publish Shiny content online one needs to create an account on shinyapps.io. The web upload within R is set up using the rsconnect package (Allaire, 2016), which needs to be installed from the command line[2]. Using a token and secret code associated with the shinyapps.io account, access from R to the account is enabled. Once the initial setup is finished, the application can be deployed in one click or with one line of code from the command line. With each upload, shinyapps.io replicates the environment from the local machine including all loaded R packages. The only requirement is that these packages need to run on Ubuntu Linux and originate from CRAN or GitHub (both from public or private repositories).

The degree of customization and performance tuning of the online app depends on the pricing plan chosen. The main restrictions to the free and low price plans are the number of applications that can be deployed, the hours per month these applications are active and the performance. Also the option to limit access to the app using user authentication is restricted to some of the higher priced plans. In this project we explored the possibilities of the free plan while developing the app and then decided to upgrade once the app got promoted within the EU FP7 EUPORIAS project and more users were accessing it.

4.3 Seasonal forecast skill app

Based on the Shiny for R package, we developed an interactive web application to explore seasonal forecast skill. This application was deployed online using the shinyapps.io service. The app creates

2 Run install.packages(„rsconnect“)

skill maps on-the-fly using pre-computed datasets of verification metrics. Thus new maps are not loaded from an archive of prepared graphics but plotted new each time the user requests them and discarded as soon as a new request is made. The app is publicly available through the following web address: https://meteoswiss-climate.shinyapps.io/skill_metrics/.

A screenshot showing the main interface of the app is presented in Figure 19. In the navigation bar on top, there are three main tabs the user can navigate and two clickable logos for EU FP7 EU-PORIAS and MeteoSwiss that link to the corresponding websites. The main part of the web application consists of a sidebar to the left, containing the elements the user can interact with (so-called widgets) and the forecast scores are displayed in the main frame in the center-right. Before users get to see the skill maps as shown in Figure 19, a welcome page is displayed. The welcome page contains some guidelines on how to use the app and can be also accessed at a later stage by clicking on the info-button just right to the title in the navigation bar.

The widgets in the sidebar to the left allow the user to select the forecast verification metrics to display. When the tab is viewed for the first time, default values are used for the input objects and the plot in the center of the screen shows a global map of the Correlation coefficient for summer (JJA) mean temperature. Three drop-down menus enable the user to set the variable (temperature or precipitation), temporal resolution and verification metric displayed. The six metrics described in the data and methods section on page 3 can be displayed for forecasts of monthly or seasonal averages. Below the drop-down menus, there is a yes/no pair of radio buttons to hide or show grid cells over the oceans (and large lakes). The two sliders at the bottom of the sidebar are used to select target month or target season and lead time to display. The names on the target slider and the length of the lead time slider depend on the temporal resolution set in the drop-down menus. In addition, the names on the lead time slider change with the target month or season such that only the available forecast initialization dates for the chosen target month (season) are shown.

When changing between forecast variables and skill metrics the lead time and target chosen are preserved. However when changing the temporal resolution, both sliders are reset to default values. This is necessary because the sliders need to be recreated and no corresponding values for the previous selection may exist. In theory it would be possible to save the value set by the user but in doing so, impatient or incautious handling of the sliders can cause the selection to jump back and forth. We decided to minimize the risk that the app will hang and therefore the sliders are rebuilt from scratch when needed despite the drawbacks. For the same reason the user will have to reselect the desired lead time when changing target months and it is not possible to select for example target month October with lead time 4 months and then go directly to target month November with the same lead time.

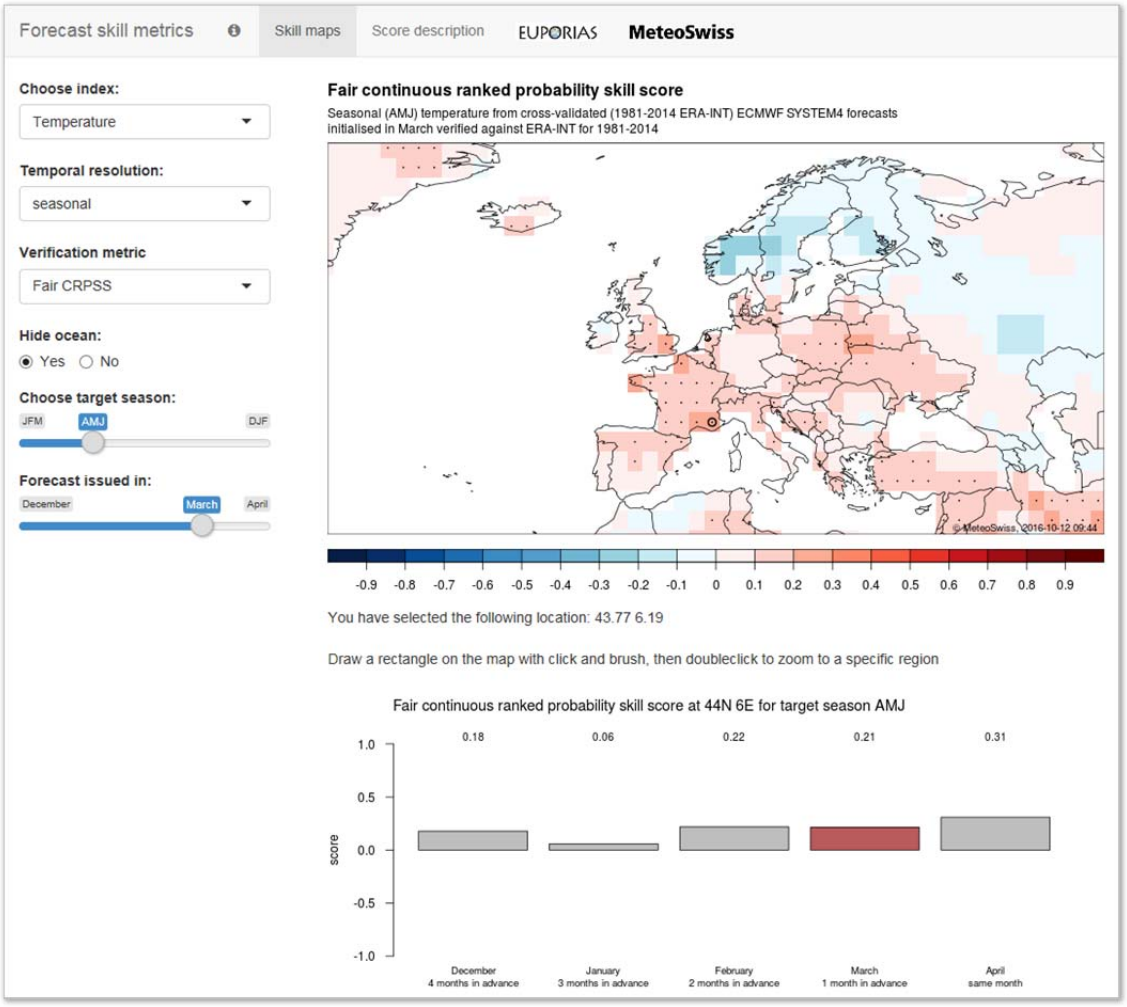

Figure 19: Screenshot of the seasonal forecast skill web application accessible on https://meteoswiss-climate.shinyapps.io/skill_metrics/. Shown is the Fair CRPSS for a zoom over Europe with a detailed view on evolution of skill with lead time for a grid point at the Côte d'Azur.

The user is not limited to the choices offered in the sidebar to explore seasonal forecast skill. It is possible to zoom and click on the map to view a specific region in detail and examine the evolution of skill with lead time for a selected grid point and target season. A single click on the map selects the center of the closest grid point and a bar plot is displayed below the map. Five or seven bars show the forecast skill at this grid point for all initialization times associated with the target season or target month selected. The current lead time displayed in the map is highlighted in red.

Using click and drag to draw a rectangle on the map, the user can select a region and a subsequent double-click will zoom to this region. Various aspects of the plot are adjusted to the selected zoom level. The significance stippling scales with the zoom level and country boundaries are displayed for sub-continental areas. Also, the aspect ratio of the map is adjusted for regions far from the equator to account for the convergence of meridians towards the poles. As the skill analysis has a resolution of 2°, users are prevented to zoom in to the grid-box level by buffering the selection to a minimum size of 24 degrees. Additional grid points are also shown around the selected area if the selection does

not match the proportions of the plot window. The zoom region is preserved when changing the input in the widgets. A double-click on the map exits the zoom view and brings the user back to the global map.

The third tab contains a concise overview of the verification metrics along with a documentation of the data and post-processing used. If a verification metric is selected in the main tab, the corresponding description subpage is loaded and displayed automatically once the user switches to the description tab. Once the description tab is selected, the user can navigate freely between the subpages to read about the data and post-processing and other metrics. When returning to the skill maps, all choices set there are conserved including zoom and grid point selection. The subpage viewed last is displayed if the selection is not changed between switching tabs back and forth.

5 Conclusions and Outlook

5.1 Prediction skill of seasonal forecasts

In section 3, we document some of the aspects of space-time variability of forecast skill by the operational ECMWF System 4 model. Skill generally decreases for longer lead times, but there are also regions where skill depends more strongly on the target month or season, rather than on lead time. Examples for the first case can be found nearly everywhere for temperature and precipitation forecasts. One example for the second case is shown in Figure 5 for temperature forecasts in a region in the Pacific influenced by ENSO and for a grid point in the same area in Figure 6. Examples for dependence on target season (or month) can be found for precipitation forecasts in Indonesia, where we find a peak in skill shortly before the monsoon season (September to November) and low predictability after the onset (December to January). Skill remains high with increasing lead time during SON and there is low predictability during DJF for all lead times. This is illustrated for a grid point in the region in Figure 10.

The verification results are well in line with the published literature on seasonal climate variability, predictability and forecast performance. The maps shown match and complement findings from other studies very well as for example for precipitation in the Indo-Pacific.

Information about skill of temperature and precipitation forecasts can be used to estimate expected skill of forecasts of indices (Bhend et al., 2016). We exemplarily show this for winter forecasts in Europe, where we assess the accuracy, reliability and discrimination of temperature and precipitation forecasts and estimate the skill of snowfall predictions. We show that there is low potential predictability of precipitation and that discrimination of temperature forecasts is not strong. From that we conclude that forecasts of snowfall for the Christmas holidays issued weeks to months ahead cannot be expected to be skillful.

We also compare the performance of forecasts for European winter with the most recent forecasts from the operational ECMWF System 4 model for winter 2016. In accordance to our results that indicate no skill for winter precipitation, the trend seen in the operational raw temperature forecasts is removed by the recalibration in place at MeteoSwiss. The verification results presented here show that there is discrimination for forecasts of temperature, however it is not strong. Furthermore overdispersion is identified as one of the main issues with European winter temperature forecasts. This indicates that the signal seen in the raw forecasts is rather an artifact of overdispersive ensembles than real discrimination of outcomes and contributes to the current scientific discussion on the possible predictability of European winters through the NAO (Scaife et al., 2014; Weisheimer et al., 2016). We can envision many other examples where the skill assessment can be used to interpret forecasts from the operational forecast models.

5.2 Experiences with Shiny and hosting Shiny apps

Developing the seasonal forecast skill application was facilitated by the R Shiny package and the public hosting on the shinyapps.io servers. Thanks to these tools, we were able to develop the app, put it online and promote it successfully in the EUPORIAS community in such short amount of time.

Shiny package for R

The Shiny package for R has proved to be powerful yet easy to learn and provides many options to create sophisticated interactive web-applications. It is suitable for a range of use cases from proof-of-concept applications for those that quickly want to build a prototype and browse through some data to full-fledged interactive applications for those who want to share a big collection of results with others and invest more in customizing appearance and interactivity options. User-friendly applications with professional appearance can be created with little effort, as the default design of Shiny is very neat and appealing. Programming of the Shiny scripts is efficient with bits of code re-executed only if required. A helpful approach for economical use of computing resources is to introduce reactive expressions and reactive values that allow arbitrary code to be run only when an event occurs that changes the reactive value.

The Shiny R package is particularly appealing to experienced R users, as knowledge about web programming languages is not required to create Shiny apps. Additional CSS and HTML can easily be added to style the app after specific visual requirements or to add functionality not supplied directly within the Shiny R package. In the seasonal forecast skill app, we use additional HTML and CSS to make small improvements to the appearance such as interactively changing the slider labels to increase usability and improve the design.

Hosting with shinyapps.io

The shinyapps.io hosting service by RStudio is quickly set up and convenient to use. Once R and the account on shinyapps.io are connected, publishing websites is straightforward. The applications work just as they do on local machines since the environment is replicated on the cloud servers. There are almost no restrictions on R packages that can be used and shinyapps.io supports old versions of R just as the newest release.

The only difference or drawbacks of the online application compared to the local version is related to performance and the time required to start up the app in particular. The seasonal forecast skill app accesses rather large data files that are uploaded together with the scripts. These files are between 2 and 12MB of size and amount to about 160MB in total. When the application is opened for the first time a new server instance is started. Before the user can use the app or indeed before she is able to see anything on the web site, all data needs to be loaded within this instance, which in our case causes a noticeable delay. Once the data is loaded, the app runs fast, no matter which inputs are changed. Additional users that enter the running server instance could in principle profit, as they will not have to wait for the data to be loaded. For our app this is probably only rarely the case, as we have set a fairly short timeout for idle instances to prevent idle instances from unnecessarily using up active hours.

The disadvantage of hosting apps with shinyapps.io is that many options are not available to users of free accounts and active hours are limited to 25 hours per month. Paying users can enlarge the in-

stance size, which could also resolve the issue with start-up speed. In the end we found the short waiting time during start-up not to be bothering and the free account was serving all our needs while developing and testing the application in-house. The upgrade to the entry-level pricing plan was necessary only after the app got promoted in the EUPORIAS community and 25 active hours per month started to become insufficient.

An alternative to using shinyapps.io would be to use a dedicated Shiny server. This obviously requires in-house knowledge, time and effort for set-up and maintenance, but running a dedicated Shiny server allows more flexibility and options to customize the server for best performance of the app. It might be worthwhile to assess this possibility for future applications, especially if it is required that data is kept in-house.

5.3 Outlook

The skill verification presented here covers several aspects of variability in forecast quality for the two forecast variables temperature and precipitation. It is a comprehensive analysis for all lead times and initializations for monthly and seasonal (3-monthly) forecasts for the ECMWF System 4. To our knowledge, this is even the most comprehensive assessment of an operational seasonal forecasting system publicly available so far. As this is a first step, a range of future extensions is conceivable.

At present, the web application is only used to display pre-computed verification metrics. To allow for more targeted applications including the verification of user-defined regional averages or indices, some of the steps performed off-line could be included in the app and performed on-the-fly. Data input and output, post-processing and the computation of the scores, however, is quite time-consuming and users are not likely to wait for minutes or hours until results become available. On-the-fly verification may therefore be difficult to implement. One way to offer the possibility to verify regional averages and different time periods would be to pre-compute the scores for each forecast and average skill scores on-the-fly. Other possible extensions include the verification for different temporal resolutions (like weakly averages). Also, the seasonal forecasting system originally has a horizontal grid spacing of about 0.7° (Molteni et al., 2011). Verification could be performed on the original grid, which would add more regional detail to the analysis.

The seasonal forecast skill app in principle works with any dataset that is brought into the same structure like the one we provide. If in the future an updated forecast validation is available, the data files can be easily replaced with the results of the new skill assessment. Also the skill of another seasonal forecast model or the verification against a different reference could be displayed using the same source code without much effort. Similar applications could also be designed to work with data from weather forecasts.

List of figures

List of tables

References

Allaire, J. J., 2016: rsconnect: Deployment Interface for R Markdown Documents and Shiny Applications. *R package* **version 0.4.3.**

Bhend, J., Mahlstein, I. and M. A. Liniger, 2016: Predictive skill of climate indices compared to mean quantities in seasonal forecasts. *Q.J.R. Meteorol. Soc*. DOI:10.1002/qj.2908

Buizza, R. and M. Leutbecher, 2015: The forecast skill horizon. *Q. J. R. Meteorol. Soc.* **141(639):** 3366-3382. DOI: 10.1002/qj.2619.

Chen, D., Zebiak, S. E., Busalacchi, A. J. and M. A. Cane, 1995: An improved procedure for El Niño forecasting: Implications for predictability. *Science.* **269:** 1699-1702.

Cornes, R. and P. Jones, 2014: Assessing the reliability of trends in extremes of surface temperature across Europe in the ERA-Interim reanalysis dataset. *EGU General Assembly Conference Abstracts.* **16.** http://adsabs.harvard.edu/abs/2014EGUGA..1615302C.

Dee, D. P., S. M. Uppala, A. J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A. C. M. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A. J. Geer, L. Haimberger, S. B. Healy, H. Hersbach, E. V. Hólm, L. Isaksen, P. Kållberg, M. Köhler, M. Matricardi, A. P. McNally, B. M. Monge-Sanz, J.-J. Morcrette, B.-K. Park, C. Peubey. P. de Rosnay, C. Tavolato, J.-N. Thépaut, and F. Vitart, 2011: The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.* **137:** 553-597.

Dee, D. & National Center for Atmospheric Research Staff (Eds), 2012: The Climate Data Guide: ERA-Interim. Retrieved from https://climatedataguide.ucar.edu/climate-data/era-interim. Last modified 11 Nov 2016 (accessed 20.11.16).

Dierks, T. and E. Rescorla, 2008: The Transpor Layer Security (TLS) Protocol**. Version 1.2.** https://tools.ietf.org/pdf/rfc5246.pdf (accessed 11.10.16).

Duan, W. and C. Wei, 2013: The 'spring predictability barrier' for ENSO predictions and its possible mechanism: results from a fully coupled model*. Int. J. Climatol.* **33:** 1280-1292. DOI: 10.1002/joc.3513

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.* **8:** 985-987.

Ferro, C. A. T., D. S. Richardson and A. P. Weigel, 2008: On the effect of ensemble size on the discrete and ranked probability scores. *Meteorol. Appl.* **15:** 19-28.

Ferro, C. A. T., 2014: Fair scores for ensemble forecasts. *Q. J. R. Meteorol. Soc.* **140:** 1917-1923. DOI: 10.1002/qj.2270.

Harvey, L. O., K. R. Hammond, C. M. Lusk, and E. F. Mross, 1992: The application of signal detection theory to weather forecasting behavior. *Mon. Wea. Rev.* **120:** 863-883.

Hastenrath S., 1987: Predictability of Java Monsoon Rainfall Anomalies: A Case Study. *J. Climate Appl. Meteor.* **26:** 133-141.

Haylock, M. and J. McBride, 2001: Spatial Coherence and Predictability of Indonesian Wet Season Rainfall. *J. Clim.* **14:** 3882-3887.

Hendon, H. H., 2003: Indonesian Rainfall Variability: Impacts of ENSO and Local Air-Sea Interaction. *J. Climate,* **16:** 1775-1790.

Hersbach, H., 2000: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Wea. Forecasting,* **15:** 559-570.

Ho, C. K., E. Hawkins, L. Shaffrey, J. Brocker, L. Hermanson, J. M. Murphy, D. M. Smith and R. Eade, 2013: Examining reliability of seasonal to decadal sea surface temperature forecasts: The role of ensemble dispersion. *Geophys. Res. Lett.* **40:** 5770-5775. DOI:10.1002/2013GL057630.

Hoerling, M. P. and A. Kumar, 2002: Atmospheric Response Patterns Associated with Tropical Forcing. *J. Clim.* **8:** 474-495.

Jha, B., A. Kumar and Z.-Z. Hu, 2016: An update on the estimate of predictability of seasonal mean atmospheric variability using North American Multi-Model Ensemble. *Clim Dyn.* DOI:10.1007/s00382-016-3217-1.

Juneng, L. and F. T. Tangang, 2005: Evolution of ENSO-related rainfall anomalies in Southeast Asia region and its relationship with atmosphere-ocean variations in Indo-Pacific sector. *Clim Dyn.* **25:** 337-350, DOI: 10.1007/s00382-0052-0031-6.

Matheson, J. E. and R. L. Winkler, 1976: Scoring rules for continuous probability distributions. *Manage. Sci.* **22:** 1087-1095.

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

Mason, S. J. and N. E. Graham, 1999: Conditional probabilities, relative operating characteristics and relative operating levels. *Weather and Forecasting.* **14:** 713-725

Mason, S. J. and A. P. Weigel, 2009: A generic forecast verification framework for administrative purposes. *Monthly Weather Review.* **137(1):** 331-349. DOI: 10.1175/2008MWR2553.1.

McPhaden, M. J., 2003: Tropical Pacific Ocean heat content variations and ENSO persistence barriers. *Geophys. Res. Lett.* **30(9),** 1480, DOI: 10.1029/2993GL016872.

Molteni, F., T. Stockdale, M. Balmaseda, G. Balsamo, R. Buizza, L. Ferranti, L. Magnusson, K. Mogensen, T. Palmer and F. Vitart, 2011: The new ECMWF seasonal forecasts system (System 4). *Technical memorandum.* **656** (unpublished).

Moron, V., A. W. Robertson and R. Boer, 2008: Spatial Coherence and Seasonal Predictability of Monsoon Onset over Indonesia. *J. Clim.* **22:** 840-850.

Mu, M., Duan, W. S. and B. Wang, 2007: Season-dependent dynamics of nonlinear optimal error growth and El Niño-Southern Oscillation predictability in a theoretical model. *J. Geophys. Res.* **112:** D10113. DOI: 10.1029/2005JD006981.

Müller, W. A., C. Appenzeller, F. J. Doblas-Reyes and M. A. Liniger, 2005a: A Debiased Ranked Probability Skill Score to Evaluate Probabilitstic Ensemble Forecasts with Small Ensemble Sizes. *J. Climat*e, **18**: 1513-1523, DOI: 10.1175/JCLI3361.1.

Müller, W., Appenzeller, C. and C. Schär, 2005b: Probabilistic seasonal prediction of the winter North Atlantic Oscillation and its impact on near surface temperature. *Clim. Dyn.* **24:** 213-226. DOI:10.1007/s00382-004-0492-z

Murphy, A. H., 1969: On the "ranked probability score". *J. Appl. Meteor.,* **8:** 988-989.

Murphy, A. H., 1971: A note on the ranked probability score. *J. Appl. Meteor.,* **10:** 155-156.

Murphy, A. H., 1993: What is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. *Wea. Forecasting,* **8:** 281-293.

Nicholls, N., 1981: Air-Sea Interaction and the Possibility of Long-Range Weather Prediction in the Indonesian Archipelago. *Mon. Wea. Rev.,* **109:** 2435-2443.

Peterson, W. W. and T. G. Birdsall, 1953: The theory of signal detectability: Part I. The general theory. *Electronic Defense Group, Technical Report* **13.** June 1953. Available from EECS Systems Office University of Michigan, 1301 Beal Avenue, Ann Arbor, MI 48109-2122 USA. Retrieved on: http://hdl.handle.net/2027.42/7068 (accessed 16.10.2016).

Robertson, A. W., V. Moron and Y. Swarinto, 2008: Seasonal predictability of daily rainfall statistics over Indramayu district, Indonesia. *Int. J. Climatol.* **29(10):** 1449-1462. DOI: 10.1002/joc.1816.

Scaife, A. A., A. Arribas, E. Blockley, A. Brookshaw, R. T. Clark, N. Dunstone, R. Eade, D. Fereday, C. K. Folland, M. Gordon, L. Hermanson, J. R. Knight, D. J. Lea, C. MacLachlan, A. Maidens, M. Martin, A. K. Peterson, D. Smith, M. Vellinga, E. Wallace, J. Waters and A. Williams, 2014: Skillful long-range prediction of European and North American winters. *Geophys. Res. Lett.* **41:** 2514-2519. DOI: 10.1002/2014GL059637.

Siegert, S., 2015. SpecsVerification: Forecast Verification Routines for the SPECS FP7 Project. R package version 0.4-1, https://CRAN.R-project.org/package=SpecsVerification

Shi, W., N. Schaller, D. MacLeod, T. N. Palmer and A. Weisheimer, 2015: Impact of hindcast length on estimates of seasonal climate predictability, *Geophys. Res. Lett.* **42:** 1554–1559. DOI:10.1002/2014GL062829.

Serreze M. C. and R. G. Barry, 2005: The Arctic Climate System. Chapter 11 p.291-334. *Cambridge University Press.* ISBN: 978-0-521-81418-8.

Shukla, J., J. Anderson, D. Baumhefner, C. Brankovic, Y. Chang, E. Kalnay, L. Marx, T. Palmer, D. Paolino, J. Ploshay, S. Schubert, D. Straus, M. Suarez and J. Tribbia, 2000: Dynamical Seasonal Prediction. *Bull Am Meteorol Soc.* **81**: 2593-2606.

Smith, D. M., A. A. Scaife, R. Eade and J. R. Knight, 2016: Seasonal to decadal prediction of winter North Atlantic Oscillation: emerging capability and future prospects. *Q. J. R. Meteorol. Soc.* **142:** 611-617. DOI: 10.1002/qj.2479.

Tanaka, M., 1994 : The Onset and Retreat Dates of the Austral Summer Monsoon over Indonesia, Australia and New Guinea. *J. Meteorol. Soc. Jap.* **72(2)** 255-267.

Tangang, F. T. and L. Juneng, 2004: Mechanisms of Malaysian Rainfall Anomalies. *J. Clim.* **17:** 3616-3622.

Taylor, A. L., S. Dessai, W. Bruine de Bruin, 2015: Communicating uncertainty in seasonal and interannual climate forecasts in Europe. *Phil. Trans.R. Soc. A* **373:** 20140454. DOI: 10.1098/rsta.2014.0454.

Torrence, C. and P. J. Webster, 1998: The annual cycle of persistence in the El Niño/Southern Oscillation. *Q. J. R. Meteorol. Soc.* **124:** 1985-2004. DOI: 10.1002/qj.49712455010.

Walker, G. T. and E. W. Bliss, 1932: World weather V. *Mem. R. Meteorol. Soc.* **4:** 53-84.

Webster, P. J., 1995: The annual cycle and predictability of the tropical coupled ocean-atmosphere system*. Meteorol. Atmos. Phys.* **56:** 33-55.

Weigel, A. P., M. A. Liniger and C. Appenzeller, 2007: The discrete brier and ranked probability skill scores. *Monthly Weather Review.* **135:** 118-124.

Weigel, A. P., M. A. Liniger and C. Appenzeller, 2009: Seasonal Ensemble Forecasts: Are Recalibrated Single Models Better than Multimodels? *Monthly Weather Review*. **137:** 1460-1479. DOI:10.1175/2008MWR2773.1

Weigel, A. P. and S. J. Mason, 2011: The generalized discrimination score for ensemble forecasts. *Monthly Weather Review.* **139(9):** 3069-3074. DOI: 10.1175/MWR-D-10-05069.1.

Weigel, A. P., 2012: Ensemble forecasts, *in Forecast Verification: A Practitioner's Guide in Atmospheric Science,* **2nd ed., chap. 8**, edited by D. B. Stephenson and I. T. Jolliffe, pp. 141-166, Wiley-Blackwell, Oxford U. K.

Weisheimer, A., Schaller, N., O'Reilly, C., MacLeod, D. A. and T. Palmer, 2016: Atmospheric seasonal forecasts of the 20th Century: multi-decadal variability in predictive skill of the winter North Atlantic Oscillation (NAO) and their potential value for extreme event attribution. *Q.J.R. Meteorol. Soc.* Accepted Author Manuscript. DOI:10.1002/qj.2976

Xue, Y., Cane, M. A., Zebiak S. E. and M. B. Blumenthal, 1994: On the prediction of ENSO: a study with a low order Markov model. *Tellus* **46A:** 512-528.

Zheng, F. and J. Zhu, 2010: Spring predictability barrier of ENSO events from the perspective of an ensemble prediction system. *Global and Planetary Change.* **72(3):** 108-117

Acknowledgement

This project was part of an internship under the supervision of Jonas Bhend and Mark Liniger. We would like to express our thanks to MeteoSwiss and EUPORIAS for supporting this work. EUPORIAS is financed by the European Commission through the 7th Framework Programme for Research, Grant Agreement 308291.

We thank all our colleagues from MeteoSwiss and EUPORIAS who tested the seasonal forecast skill web application, gave us feedback and contributed to the improvement of the application.

Also we would like to thank Katrin Sedlmeier who acted as reviewer of this report.

A Appendix

Spring Predictability Barrier in the tropical Pacific

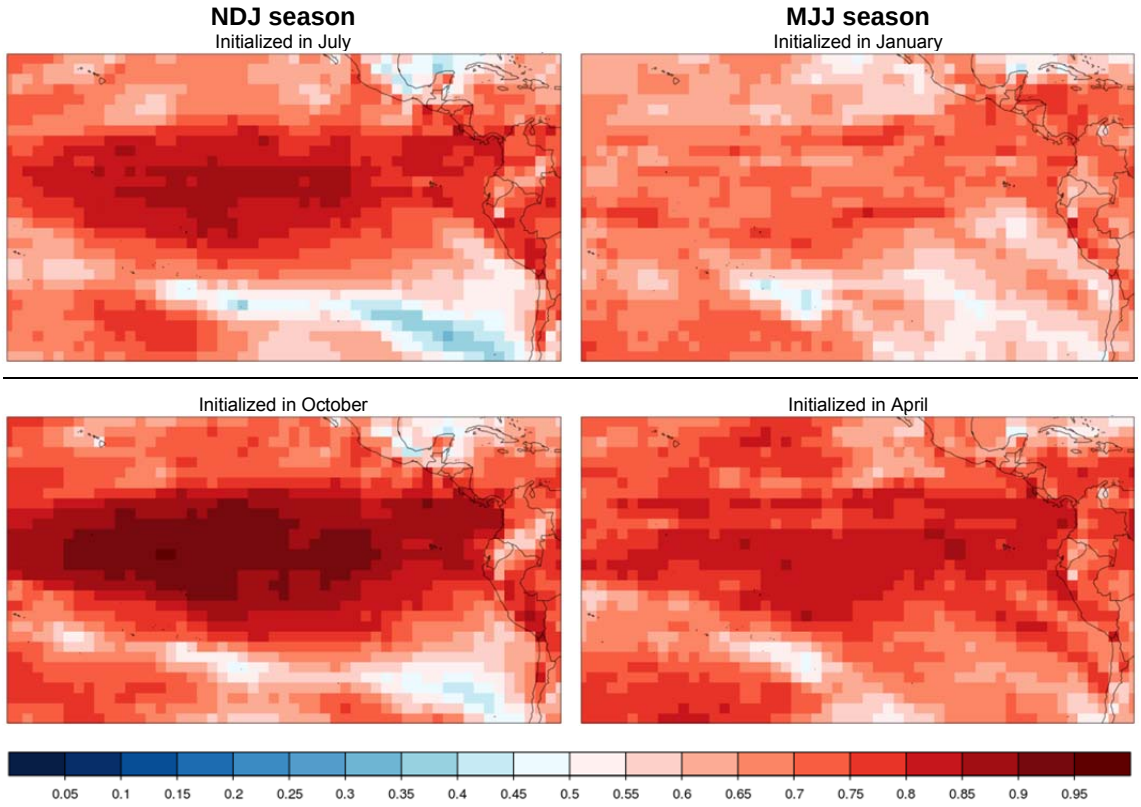

Figure A-1: Generalized Discrimination Score for seasonal temperature forecasts in the ENSO region. The plots show the area between 170 °W to 70 °W and 30 °S to 20 °N. The forecasts on the left are issued for the NDJ season and to the right forecasts for the MJJ season are shown. The lead time decreases from top to bottom so that forecasts with lead times 5 to 7 months are shown at the top and lead times 2 to 4 months at the bottom. Blueish colors indicate a score below 0.5, which means that the forecast has no discrimination and performs worse than guessing in telling different cases apart. Reddish colors indicate that the forecast has discrimination.

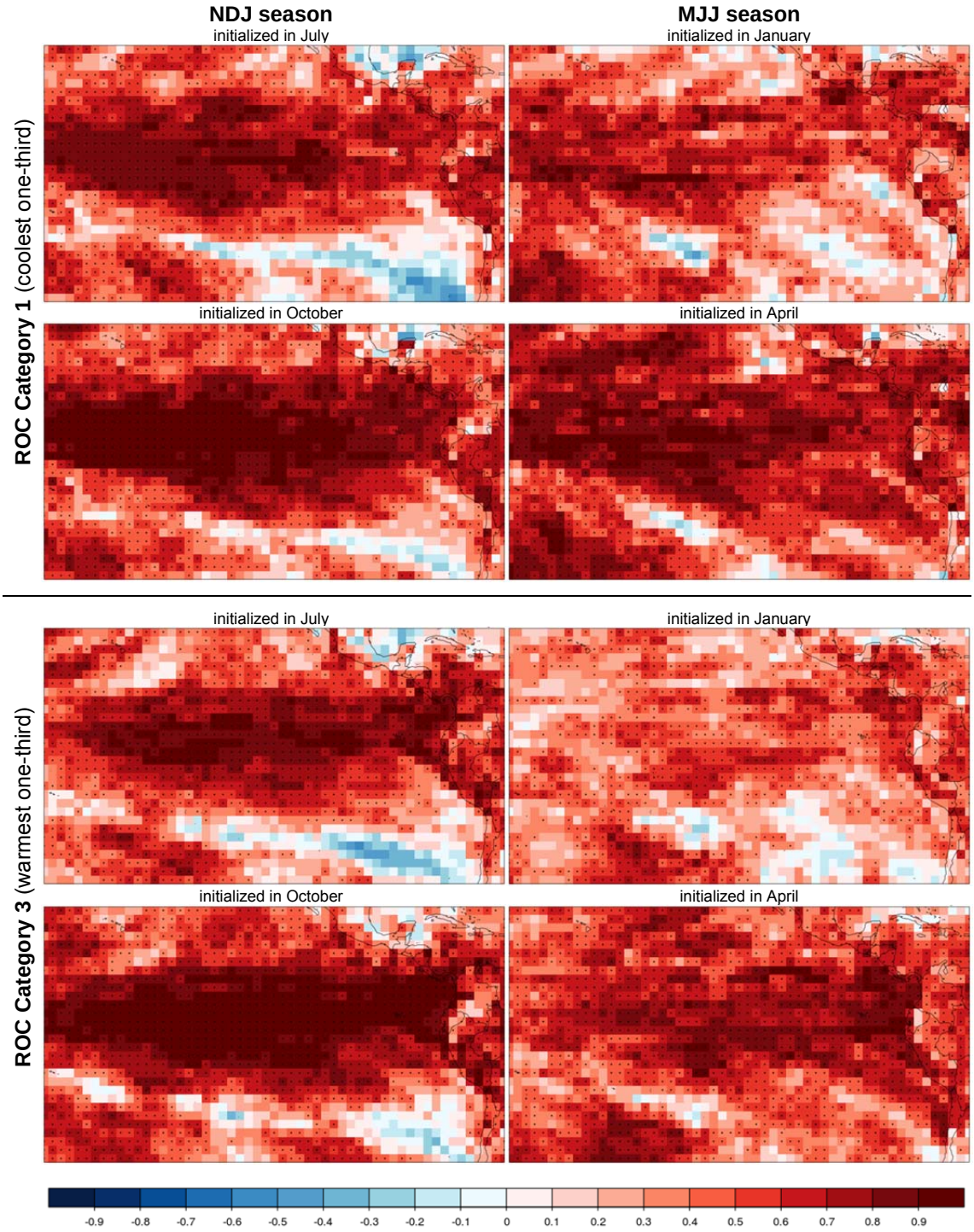

Figure A-2: ROC area score of temperature forecasts in the ENSO regions showing skill of predicting the coolest one-third (top panel of four plots) and warmest one-third (bottom panel of four plots). Forecasts for the NDJ season are shown to the left and for the MJJ season to the right. Top plots in the category 1 and the category 3 panel show forecasts with lead times 5 to 7 months and the bottom plots lead times 2 to 4 months. Positive values shown in red indicate that the forecast outperforms climatology. Forecasts significantly (at the 5% level) better than guessing the category are indicated by stippling of the respective grid cell.

Monsoon precipitation in the Indo-Pacific

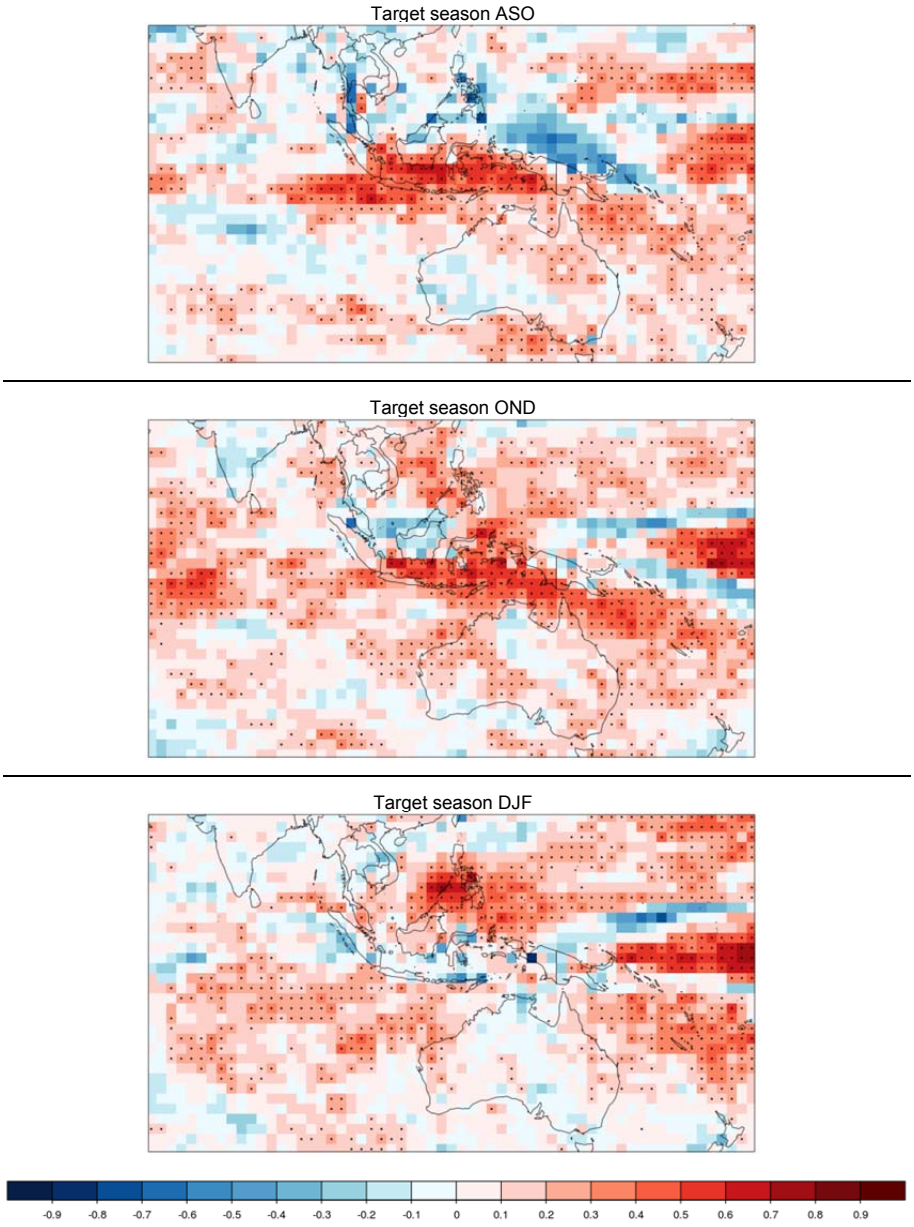

Figure A-3: Fair RPSS for seasonal precipitation forecasts in the Indo-Pacific including Australia. Forecasts for lead months 2-4 initialized in July, September and to November are shown.

MeteoSchweiz
Operation Center 1
CH-8044 Zürich-Flughafen
T +41 58 460 99 99
www.meteoschweiz.ch

MeteoSvizzera
Via ai Monti 146
CH-6605 Locarno Monti
T +41 58 460 97 77
www.meteosvizzera.ch

MétéoSuisse
7bis, av. de la Paix
CH-1211 Genève 2
T +41 58 460 98 88
www.meteosuisse.ch

MétéoSuisse
Chemin de l'Aérologie
CH-1530 Payerne
T +41 58 460 94 44
www.meteosuisse.ch