A global empirical system for probabilistic seasonal climate prediction

Eden, JM, Van Oldenborgh, GJ, Hawkins, E & Suckling, EB

Geoscientific
Model Development

A global empirical system for probabilistic seasonal climate prediction

J. M. Eden[1], G. J. van Oldenborgh[1], E. Hawkins[2], and E. B. Suckling[2]

[1]Royal Netherlands Meteorological Institute (KNMI), De Bilt, the Netherlands
[2]NCAS-Climate, Department of Meteorology, University of Reading, Reading, UK

Correspondence to: J. M. Eden (jonathan.eden@knmi.nl)

Abstract. Preparing for episodes with risks of anomalous weather a month to a year ahead is an important challenge for governments, non-governmental organisations, and private companies and is dependent on the availability of reliable forecasts. The majority of operational seasonal forecasts are made using process-based dynamical models, which are complex, computationally challenging and prone to biases. Empirical forecast approaches built on statistical models to represent physical processes offer an alternative to dynamical systems and can provide either a benchmark for comparison or independent supplementary forecasts. Here, we present a simple empirical system based on multiple linear regression for producing probabilistic forecasts of seasonal surface air temperature and precipitation across the globe. The global CO_2-equivalent concentration is taken as the primary predictor; subsequent predictors, including large-scale modes of variability in the climate system and local-scale information, are selected on the basis of their physical relationship with the predictand. The focus given to the climate change signal as a source of skill and the probabilistic nature of the forecasts produced constitute a novel approach to global empirical prediction.

Hindcasts for the period 1961–2013 are validated against observations using deterministic (correlation of seasonal means) and probabilistic (continuous rank probability skill scores) metrics. Good skill is found in many regions, particularly for surface air temperature and most notably in much of Europe during the spring and summer seasons. For precipitation, skill is generally limited to regions with known El Niño–Southern Oscillation (ENSO) teleconnections. The

system is used in a quasi-operational framework to generate empirical seasonal forecasts on a monthly basis.

1 Introduction

The provision of reliable seasonal forecasts is an important area in climate science and understanding the limitations and quantifying uncertainty remains a key challenge (Doblas-Reyes et al., 2013; Weisheimer and Palmer, 2014). Operational seasonal forecasting, although once limited to a handful of research centres, is now a regular activity across the globe. Much recent focus has been given to the skill and reliability of seasonal climate predictions. Dynamical (process-based) forecast systems are arguably the most important tool in producing predictions of seasonal climate at continental and regional scales. Such systems are based on numerical models that represent dynamical processes in the atmosphere, ocean and land surface in addition to the linear and non-linear interactions between them. However, the development of dynamical systems is a continuous challenge; climate models are inherently complex and computationally demanding and often contain considerable errors and biases that limit model skill in particular regions and seasons.

As an alternative to dynamical forecast systems, empirical approaches aim to describe a known physical relationship between regional-scale anomalies in a target variable (the predictand) – say, temperature or precipitation – and preceding climate phenomena (the predictors). In its simplest form, an empirical forecast may be based on persistence in which observations of a given variable at some lead time are

taken as the forecast for that variable. Such forecasts have frequently performed better at short lead times than those simply prescribed by the long-term climatology, particularly so in the tropics. More sophisticated statistical methods include analog forecasting (van den Dool, 2007; Suckling and Smith, 2013) and regression-based techniques, which may in turn take predictive information from spatial patterns using, for instance, empirical orthogonal functions (EOFs) (e.g. van Oldenborgh et al., 2005b), maximum covariance analysis (MCA) (e.g. Coelho et al., 2006), and linear inverse modelling (LIM) (e.g. Penland and Matrosova, 1998). Empirical predictions for the phase and strength of the El Niño–Southern Oscillation (ENSO) have historically shown comparable skill to those produced by dynamical systems (e.g. Sardeshmukh et al., 2000; Peng et al., 2000; van Oldenborgh et al., 2005b). Additionally, an inherent advantage of empirical methods is the ease with which knowledge of climate variability gained from analysis of up-to-date observations can be incorporated into a prediction system (Doblas-Reyes et al., 2013), which in turn facilitates the development of new methodologies and statistical techniques (van den Dool, 2007).

Empirical forecasts serve both as a baseline for dynamical models and can be used to improve the forecasts by limiting the effects of dynamical model biases. However, differences in the development and output of dynamical and empirical statistical approaches make systematic comparison troublesome, and understanding the relative skill of each forecast type is challenging. Recent attempts have been made in developing empirical benchmark systems for multiple variables, such as land and sea surface temperature, on decadal timescales (e.g. Ho et al., 2013; Newman, 2013), concluding that the usefulness of such systems merits further development. While comparison of dynamical and empirical systems for seasonal forecasts is not novel, a systematic global comparison for multiple variables, including probabilistic measures, has been lacking. A key potential benefit of such comparison is the identification of regions where empirical models are skilful and may be able to provide useful forecast information to complement the output of dynamical systems. Supplementing dynamical forecasts with empirical forecasts is of great importance in situations where dynamical systems are known to have weaknesses. It has also been shown that combining the output of empirical and dynamical systems can produce marked improvement over single-system forecasts (e.g. Coelho et al., 2006; Schepen et al., 2012).

A fundamental criticism of empirical systems is the question of their applicability in a future, perturbed climate. In other words, to what extent will the predictor–predictand relationships underpinning a statistical model remain stationary under climate change? Sterl et al. (2007) found that, within the statistical uncertainties, no changes could be detected in ENSO teleconnections. Doblas-Reyes et al. (2013) recently noted that the temporal evolution of seasonal climate should be considered as forced not only by the inter-

nal variability of the climate system but also by changes in concentrations of greenhouse gas and aerosols as a result of anthropogenic activities. Such external forcings are considered in climate change simulations and also to an increasing extent in the field of decadal prediction (e.g. Krueger and Von Storch, 2011). Current seasonal forecast systems now include these forcings (Doblas-Reyes et al., 2006; Liniger et al., 2007), but the resulting trends are sometimes not realistic.

Here we present and validate a simple empirical system for predicting seasonal climate across the globe. The prediction system, based on multiple linear regression, produces probabilistic forecasts for temperature and precipitation using a number of predictors based on well-understood physical relationships. The global equivalent CO_2 concentration is an indicator of the climate change signal and is used as the primary predictor in all forecasts. Additional predictors describing large-scale modes of variability in the climate system, starting with ENSO, and local-scale information are subsequently selected on the basis of their potential to provide additional predictive power. The system presented will have two purposes: (a) to serve as a benchmark for assessing and comparing the skill of dynamical forecast systems, and (b) to act as an independent forecast system in combination with predictions from dynamical systems. Key to achieving these goals will be the system's implementation in a quasi-operational framework with empirical forecasts made on a monthly basis and the availability of a set of hindcasts.

The method implemented here constitutes a relatively simple approach to empirical forecasting. The global and automated nature of the prediction system calls for the underlying empirical method to be parsimonious in terms of the predictive sources used to construct it. The statistical model and the selection of predictors will thus be based on physical principles and processes to the fullest extent so as to elicit the maximum predictive power of, first of all, the long-term trend associated with the climate change signal and, secondly, as few additional predictors as is necessary in order to minimise the risk of overfitting. The final system will also be sufficiently flexible to facilitate its future development. Such development may involve inclusion of additional predictors should more complete and reliable data sets become available, or the application of the system to alternative predictands including those relating to the magnitude and frequency of extreme events.

Producing empirical forecast output in similar format to dynamical systems is crucial when designing a framework for robust comparison. A weakness of current dynamical–empirical system comparison is the general lack of a common set of validation measures. Whereas dynamical systems inherently provide output in the form of ensemble forecasts, which may be validated in probabilistic terms, validation of empirical systems does not always extend beyond deterministic measures, such as bias, root-mean-squared error (RMSE) and correlation (Mason and Mimmack, 2002). Here,

the uncertainties are explicitly parametrised as an ensemble of forecasts and we employ a rigorous validation framework designed to assess both the deterministic and probabilistic aspects of the forecast system.

The remainder of the paper is structured as follows. Section 2 describes the prediction system in full, including the observational data used for empirical model fitting and validation. An analysis of the potential usefulness of the predictors is given in Sect. 3. The skill of the prediction system is then assessed in Sect. 4 with a discussion and outlook given in Sect. 5.

2 Prediction system outline

Key to achieving the goals set out in Sect. 1 is the development of an automated forecast system that can be applied globally and, in principle, for any number of predictands. For these reasons, the regression-based prediction system developed here is relatively simple in comparison with more sophisticated statistical models, with emphasis given to a basis of physical processes and the avoidance of overfitting.

Our system incorporates a multiple linear regression approach for estimating seasonal (3-month) surface air temperature (SAT) and precipitation (PREC) as a function of global and local atmospheric and oceanic fields. The approach used assumes the predictand time series x to consist of two components:

$$x = x^{\text{ext}} + x^{\text{int}}, \tag{1}$$

where x^{ext} is the response to externally forced low frequency variability associated with anthropogenic activity and x^{int} represents the internal variability independent of changes in external forcing (Krueger and Von Storch, 2011). We seek first to utilise the predictive information in x^{ext} which is assumed to be linearly dependent on the global CO_2-equivalent concentration (CO2EQV), based on historical estimates until 2005 and according to Representative Concentration Pathway (RCP) 4.5 thereafter, which constitutes the net forcing of greenhouse gases, aerosols, and other anthropogenic emissions (Meinshausen et al., 2011). Secondly, we seek to identify a set of predictors that best represents x^{int}. The predictand time series x may be modelled as a function of a set of predictors:

$$x = \alpha + \beta C + \sum_{i=1}^{n} (\Phi_i F_i) + \epsilon, \tag{2}$$

where C is CO2EQV at a given lead time and F is a set of n additional predictors at the same lead time that describes x^{int}. The regression parameters β and Φ are those required to transform C and F, respectively, α is the constant regression term and ϵ is the set of residuals specific to the model fit. In this case, predictors are taken from the previous 3-month season at a lead time of 1 month (e.g. the forecast for the

season March–April–May is estimated using predictors from November–December–January). An independent regression model is calibrated at each grid point. Whereas CO2EQV is included as a predictor by default, all additional predictors are included on the basis of their predictive potential, which is determined by a predictor selection procedure prior to model fitting. In the remainder of this section we (a) identify potential predictors and describe the sources of both predictor and predictand data (Sect. 2.1) and (b) provide further details on the predictor selection approach, the model fitting procedure and the validation framework (Sect. 2.2).

2.1 Potential predictors

As additional predictors F, we consider first of all variables that describe large-scale modes of variability. ENSO is the most important of these in terms of its contribution to the skill of seasonal predictions, particularly in the tropics (van Oldenborgh et al., 2005a; Balmaseda and Anderson, 2009; Weisheimer et al., 2009; Doblas-Reyes et al., 2013). Circulation and precipitation patterns in the tropical Pacific associated with ENSO sea surface temperature (SST) anomalies are subsequently linked to climate variability in other parts of the globe (Alexander et al., 2002). In addition, modes of variability in other tropical oceans, including the tropical Atlantic and Indian basins, are known to contribute substantially to variability in SAT and PREC, particularly in surrounding regions (Doblas-Reyes et al., 2013). Many such phenomena are linked in some way to ENSO, although variability in the Indian Ocean Dipole (IOD) is known to occur independently (Zhao and Hendon, 2009). Similarly, the Pacific Decadal Oscillation (PDO), defined as the leading empirical orthogonal function of North Pacific monthly SST anomalies, is considered as a representation of variability on interdecadal timescales that is not otherwise apparent in interannual ENSO variability (Liu and Alexander, 2007). Drought occurrence in the United States is known to be linked to the phase of both PDO and Atlantic Multidecadal Oscillation (AMO). Atmospheric anomalies, including troposphere–stratosphere interactions, are also known to have predictive potential. The Quasi-Biennial Oscillation (QBO) (Ebdon and Veryard, 1961; Baldwin et al., 2001) has recently been considered in a multiple regression model for predicting European winter climate (Folland et al., 2012). With this in mind, the following indices are considered as predictors: NINO3.4 (representative of ENSO), PDO, AMO, IOD and QBO. The system is designed to be flexible enough for the inclusion of additional predictors in the future.

External forcing and global modes of variability are not the only source of skill in seasonal forecasts. Many studies, including those based on dynamical systems, have found links between local climate and variations in preceding nearby climate phenomena (e.g. van den Hurk et al., 2012; Quesada et al., 2012). The most simple of these is persistence; that is, the value of the predictand (either SAT or PREC) for the

Table 1. Description of predictor variables and their sources.

Predictor	Source
CO2EQV	CO_2-equivalent concentrations (Meinshausen et al., 2011)
NINO3.4	Calculated from SST fields from HadISST (Rayner et al., 2003)
PDO	University of Washington (http://research.jisao.washington.edu/pdo/)
QBO	At 30 hPa from the reconstruction of Brönnimann et al. (2007)
AMO	Calculated by van Oldenborgh et al. (2009); based on HadSST (Kennedy et al., 2011a, b)
IOD	Calculated from SST fields from HadISST (Rayner et al., 2003)
LSST	HadSST3 (Kennedy et al., 2011a, b)
CPREC	GPCC Full Data Reanalysis version 6 (Schneider et al., 2011)

same location at some lead time. Here, we seek to elicit predictive information from persistence (PERS) and other variables that vary from grid point to grid point in addition to the set of large-scale modes of variability described above. For coastal locations in particular, we seek to maximise the potential of short-term memory contained within neighbouring sea surface temperatures to provide greater predictability than PERS at the specified lead time. We derive a local sea surface temperature (LSST) index for each predictand grid cell, defined as the mean of the k nearest grid cells containing SST information. Here, $k = 5$ throughout the analysis although this value could of course be altered or optimised for region-specific analysis. Finally, as a proxy for soil moisture, which has been shown to impact on local temperature (e.g. van den Hurk et al., 2012), we also consider accumulated rainfall (CPREC) as a potential predictor.

Further details of the sources of predictor data are given in Table 1. Our list of predictors is not exhaustive. Much recent work has sought to identify predictability arising from the extent of sea ice and snow covered land, the reflective and insulative attributes of which are relevant for SAT and PREC in several regions of the extra-tropics (e.g. Shongwe et al., 2007; Dutra et al., 2011; Brands et al., 2012; Chevallier and Salas-Mélia, 2012). However, these variables are not considered for the present system due to the absence of sufficiently long and reliable data sets, although some effects are effectively captured by persistence. The design of the prediction system facilitates inclusion of additional predictors should high quality observational or reanalysis data become available.

2.2 Model fitting and validation

Global observational data sets provide the predictand (SAT and PREC) fields required for model calibration and validation. SAT is taken from the Cowtan and Way (2014) reconstruction of the Hadley Centre–Climatic Research Unit Version 4 (HadCRUT4) (Morice et al., 2012), which uses kriging to account for missing data in unsampled regions. PREC is taken from the Global Precipitation Climatology Centre (GPCC) Full Data Reanalysis version 6 (Schneider et al., 2011) for the period 1901–2010 combined with additional data for the period 2011–2013 taken from the GPCC monitoring product following bias correction.

Analysing the degree of additional predictive skill offered by each predictor will form an important precursor to the implementation of the system. A two-step predictor selection procedure is used to determine the fewest numbers of predictors necessary to provide greatest predictive skill. The selection procedure may be considered "offline" in the sense that it is implemented prior to model fitting. In the first step, global maps of linear correlation between predictand–predictor pairs form a basis for a physical understanding of the factors governing variability. Predictors that show good potential and do not exhibit colinearity with other predictors are included in the second step: the selection of predictors to be passed to the empirical forecast model itself.

To achieve this, the linear trend associated with CO2EQV is first of all removed from both the predictand x and the set of predictors F by fitting the models

$$x = \alpha_1 + \beta_1 C + \epsilon^x \tag{3}$$

and

$$F_i = \alpha_2 + \beta_2 C + \epsilon^{F_i}, \tag{4}$$

where α_1, β_1 and α_2, β_2 are the respective regression parameters for each model fit and ϵ^x and ϵ^{F_i} are the time series of residuals that equate to the detrended predictand and predictors, respectively. Correlation is performed between ϵ^x and each of the N predictors within the set ϵ^{F_i} (where $i = 1, 2, \ldots N$). Predictors that exhibit significant (at the 90 % level) correlation are identified. The two-step approach is designed to avoid overfitting, which would lower skill scores, and to ensure that the empirical model is built on physical principles to the fullest extent. The first step is to an extent qualitative and undertaken only once for each predictand, i.e. for each predictand there is an agreed set of potential predictors independent of season or location. However, the fully quantitative second step is performed independently at each grid point and for each season. Following the selection of predictors, all significant predictors are then entered into a multiple linear regression along with CO2EQV; Eq. (2) is
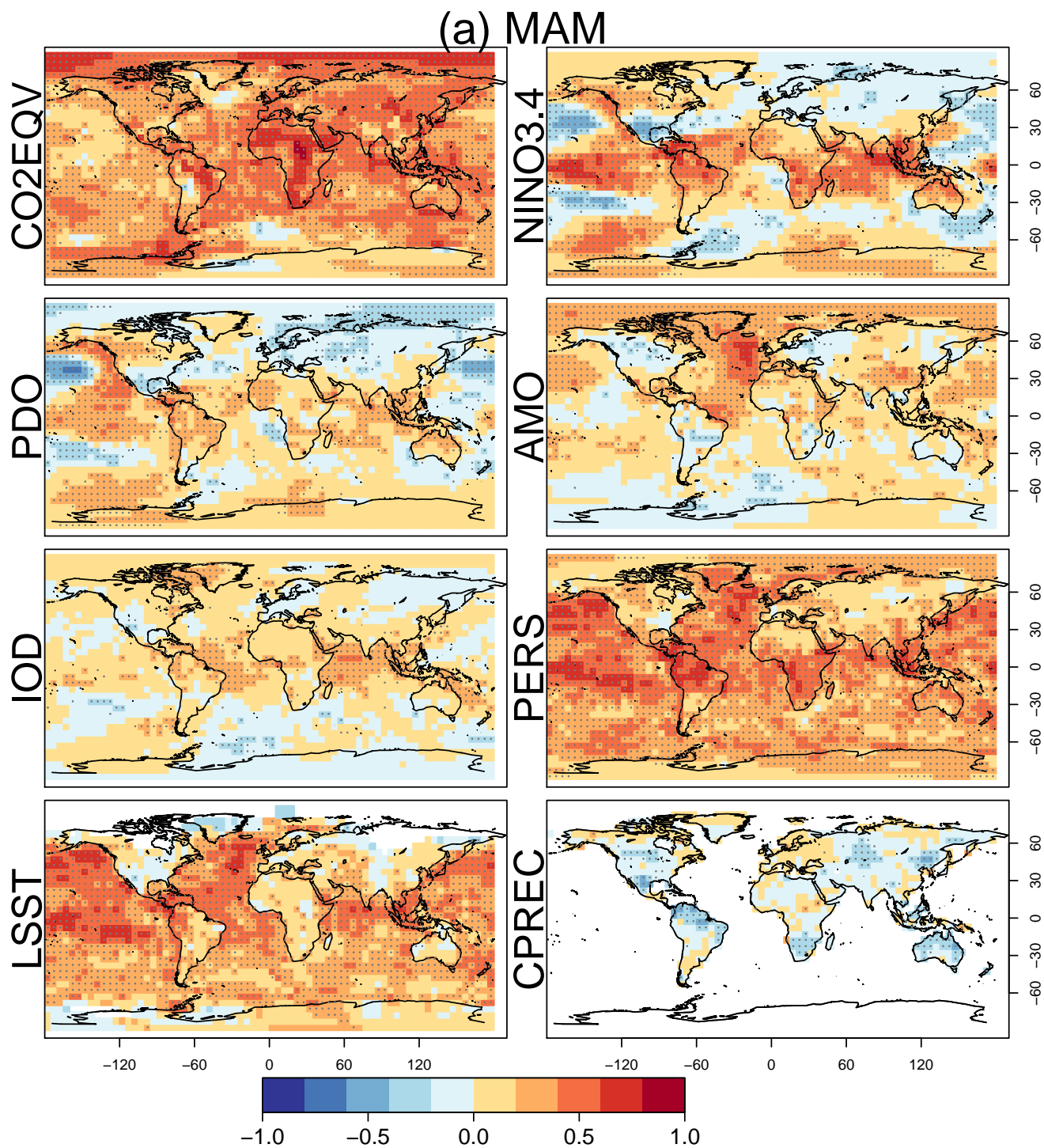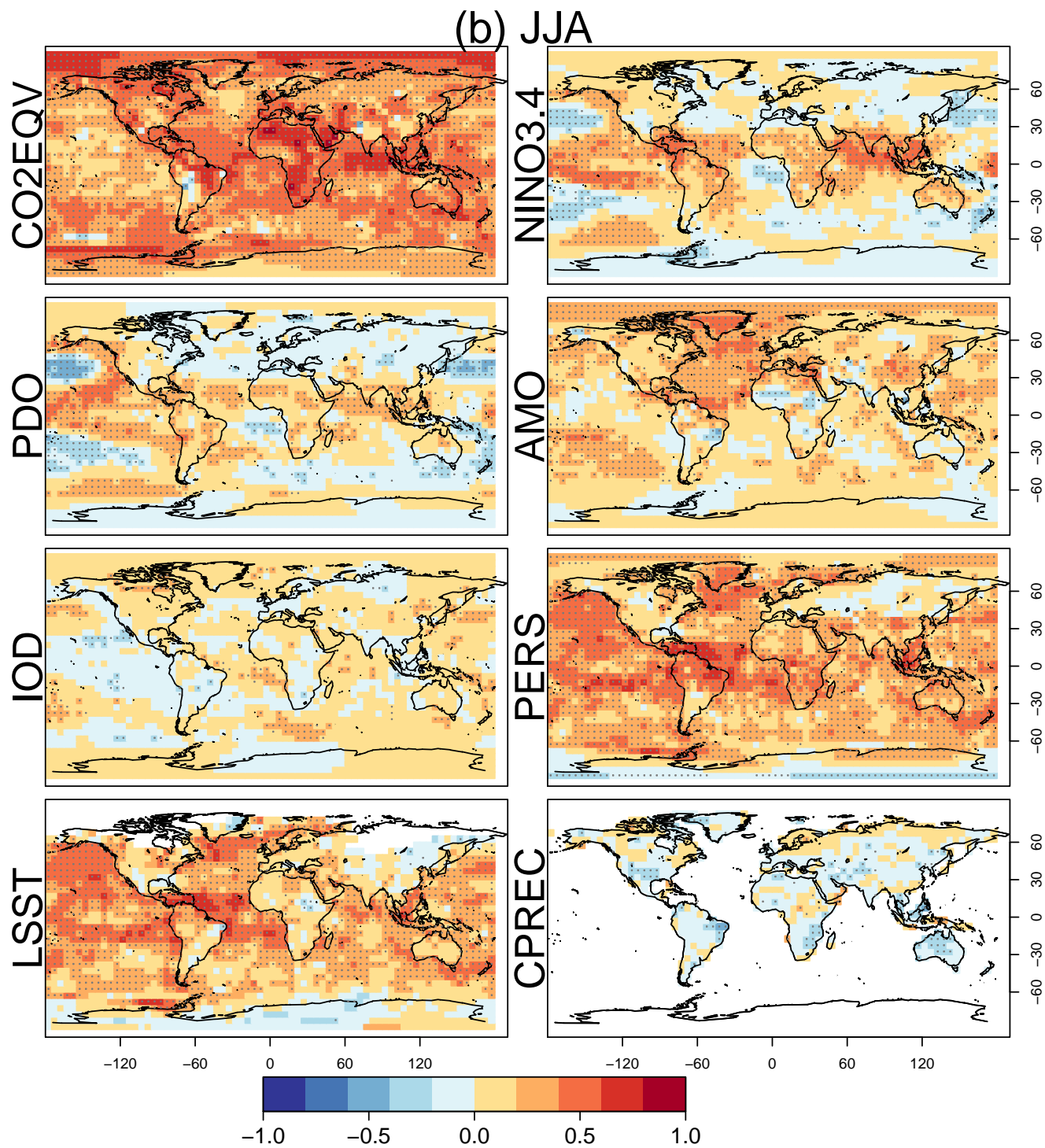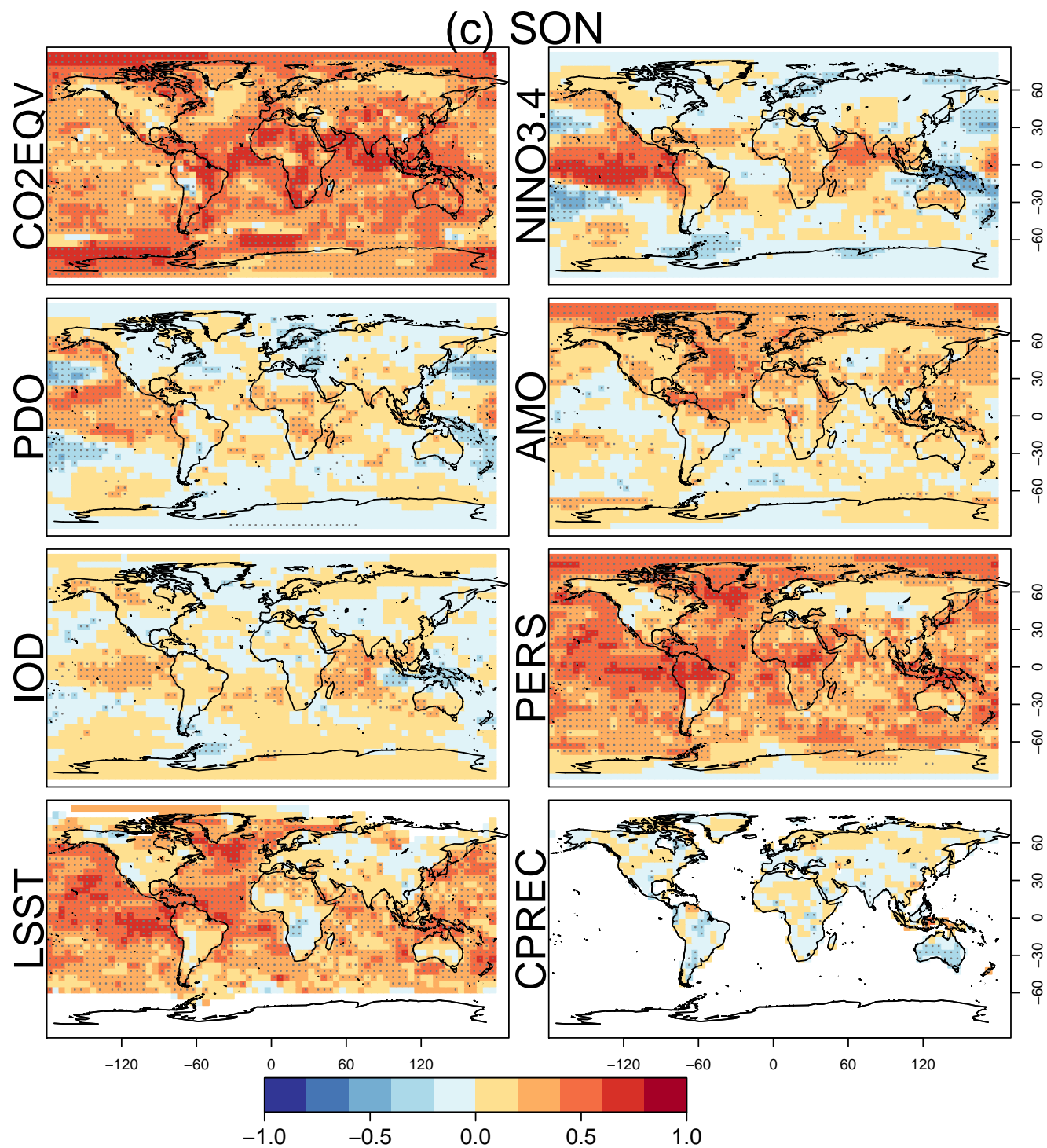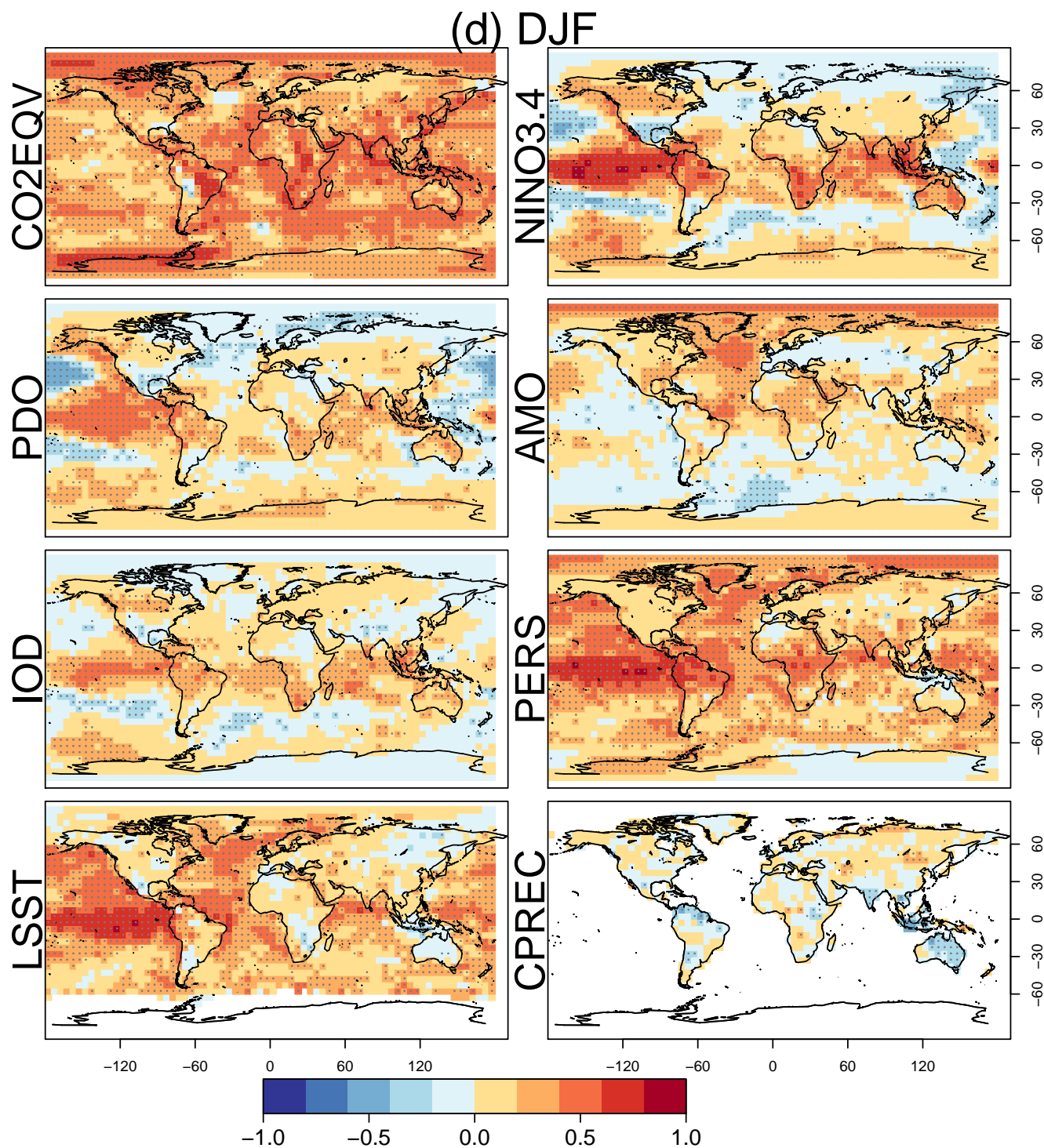
Figure 1.

Figure 1.

Figure 1.

Figure 1. Correlation between SAT and the set of predictors with a 1 month lead time (1961–2013) for **(a)** MAM, **(b)** JJA, **(c)** SON, and **(d)** DJF. Correlation between CO2EQV is shown in the top left panel; subsequent panels show correlation between predictand–predictor pairs following removal of the CO2EQV trend. Stippling is used to indicate significance at the 95 % level.
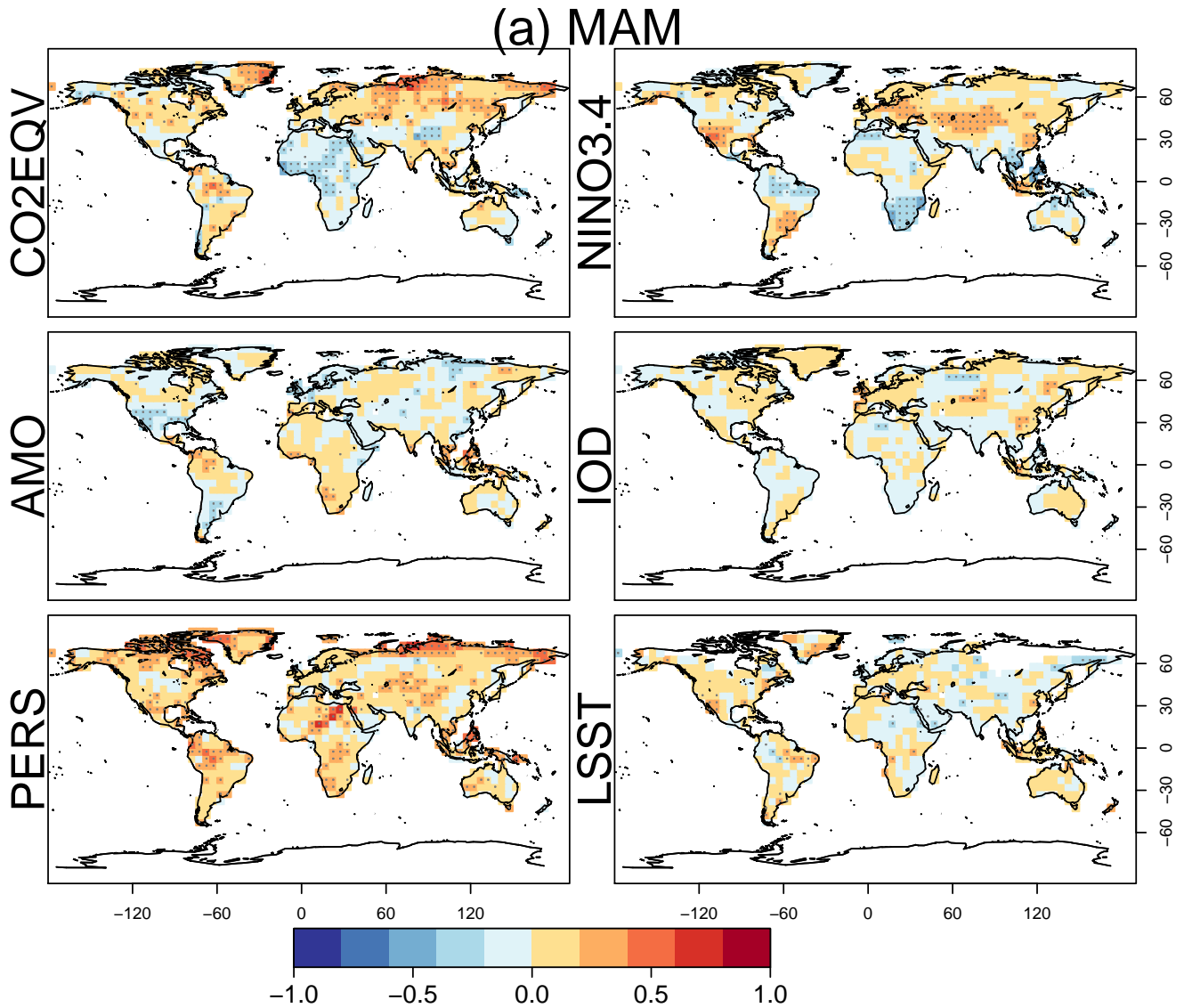
(a) MAM

Figure 2.

thus modified:

$$x = \alpha + \beta C + \sum_{i=1}^{k} (\Phi_i F_i^S) + \epsilon, \qquad (5)$$

where F^S is the subset of k predictors from F that meet the significance criteria outlined in the selection procedure. An estimate for the unknown predictand $\hat{x}$ at forecast time t may be determined:

$$\hat{x}_t = \alpha + \beta C_t + \sum_{i=1}^{k} (\Phi_i F_{i_t}^S). \qquad (6)$$

A key component of the empirical prediction system is the provision of probabilistic output. The residuals ϵ from the regression fit in Eq. (5) are randomly sampled (with replace-

ment) and subsequently used to generate a forecast ensemble. The kth member of the ensemble $\hat{x}^{ens}$ at forecast time t is thus given by

$$\hat{x}_{t,k}^{ens} = \hat{x}_t + \epsilon_k, \qquad (7)$$

where ϵ_k is a randomly sampled member of ϵ. Sampling of the residuals is performed 51 times, reflecting the typical ensemble size in an operational dynamic forecast. The ensemble allows for the calculation of probabilistic skill scores and will provide a basis for full comparison with the output of dynamical systems. It is anticipated that future development of the system will consider more complex methods of ensemble generation.

The model is calibrated and validated in a hindcast framework using a causal approach: hindcasts are produced for
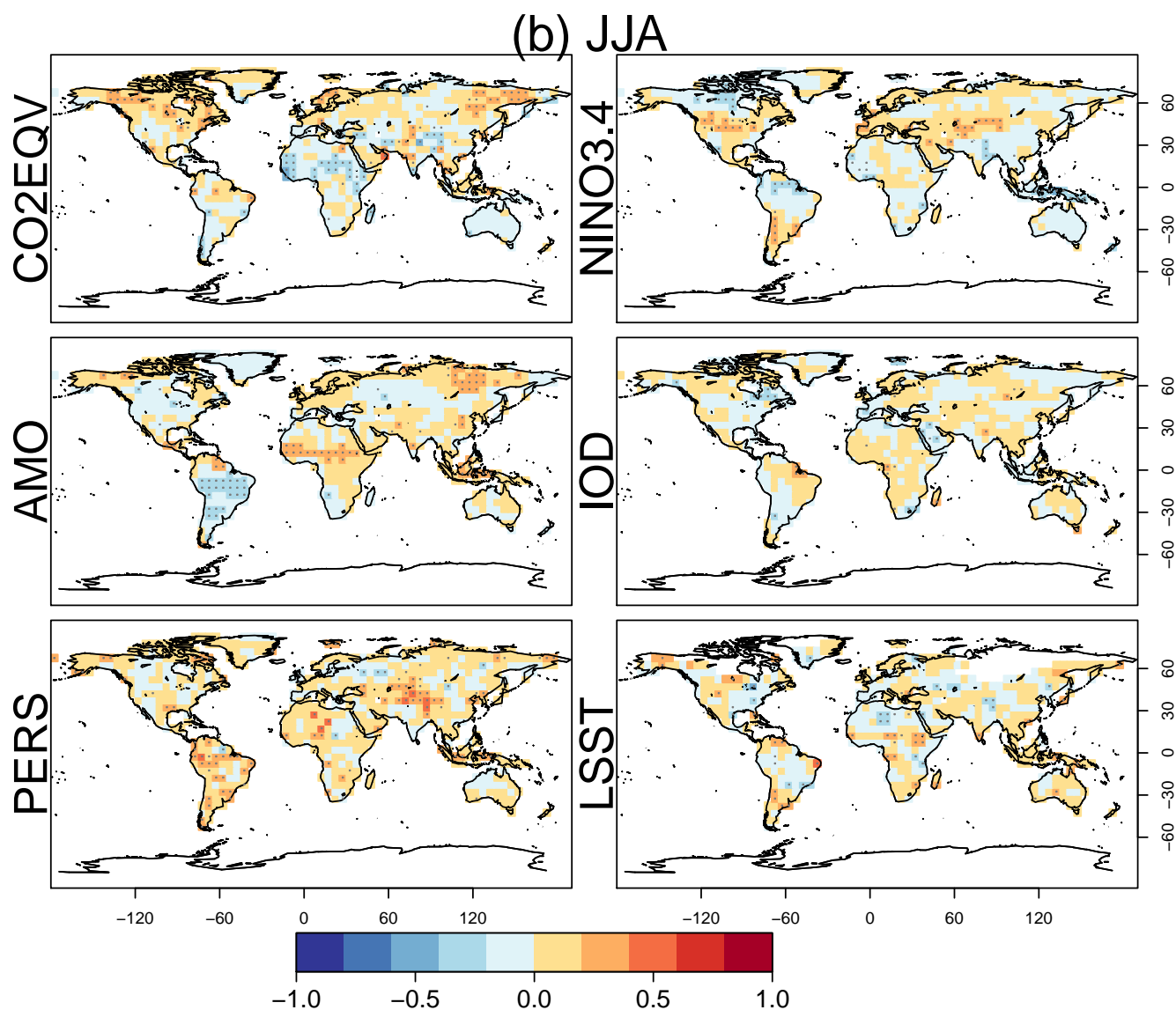
(b) JJA

Figure 2.

1961–2013 using data since 1901 prior to the hindcast start date. The causal approach was chosen instead of a leave-one-out framework in order to replicate the set of observational data that would have been available for each hindcast were it produced in real time. The predictor selection procedure, in addition to being location-specific, is also implemented independently for each hindcast. In other words, for a given grid point, a given predictor would only be included in the regression model for hindcasts with fitting periods during which it demonstrates predictive potential, allowing for the maximum value to be taken from predictor information in the fairest way. It is also important to note that, in setting the earliest hindcast to 1961, we seek to limit the impact of poor quality available predictand and predictor data in the early 20th century. Additionally, to ensure robustness, the multiple linear

regression model requires a complete predictand–predictor time series of at least 30 years in the fitting period for a forecast to be produced.

Both the deterministic and probabilistic aspects of the prediction system must be systematically validated using a number of measures. Global maps of correlation between hindcast estimates and observations provide a view on the degree of representation of temporal variability. Verification scores originally developed in the context of numerical weather prediction, including the root-mean-squared error skill score (RMSESS) and the continuous rank probability skill score (CRPSS) (e.g. Ferro, 2013), provide a quantification of the degree of bias and the skill of the probability distribution produced by the ensemble, respectively. Such verification measures are also used to determine skill scores that describe
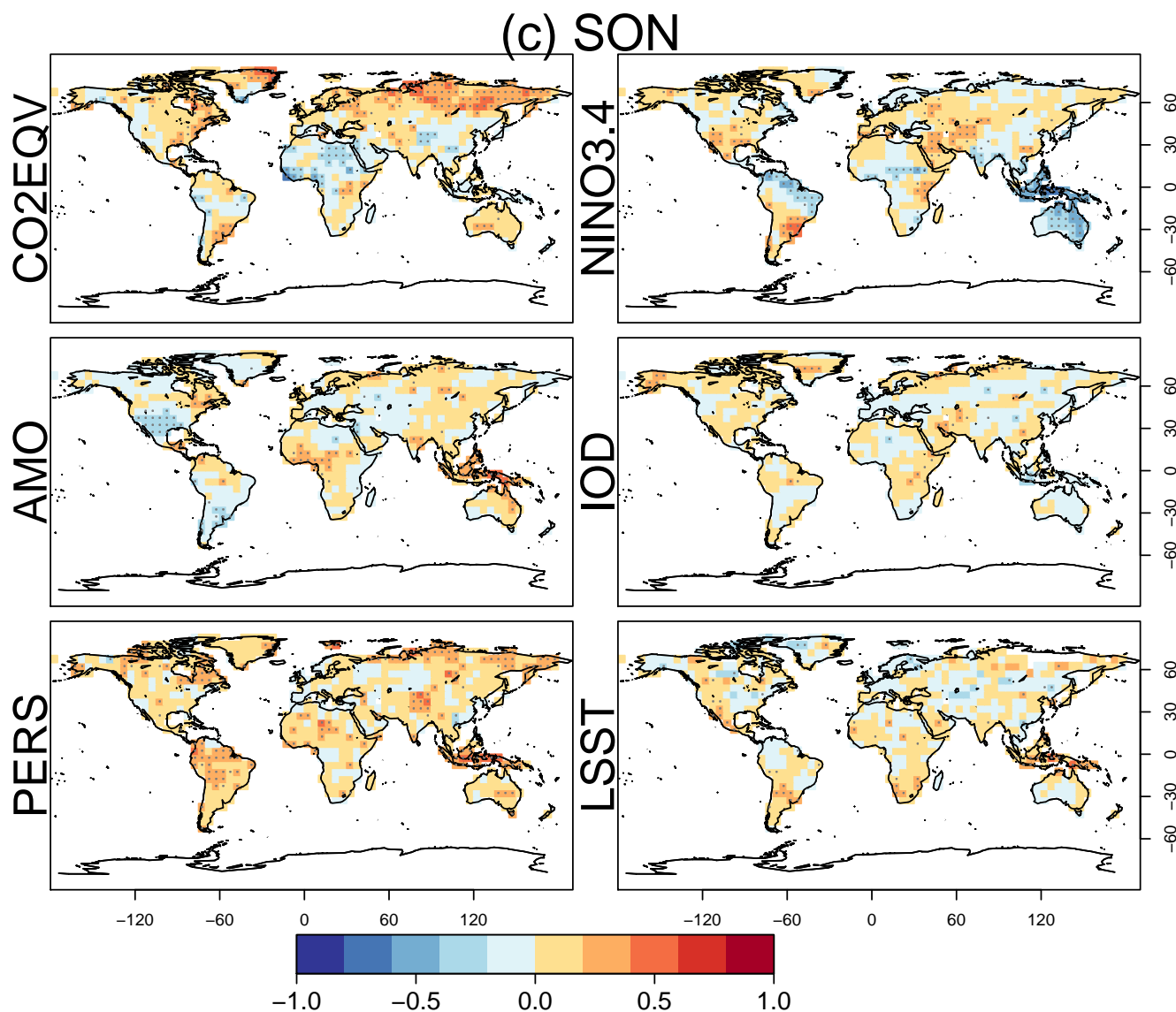
Figure 2.

forecast skill against a reference ensemble forecast. The reference forecast is produced by random sampling of the climatology, i.e. the observations for each year in the fitting period.

3 Analysis of potential predictors

3.1 Surface air temperature

The surface air temperature (SAT) shows a clear trend almost everywhere, which is assumed to be proportional to the forcing of greenhouse gases, described by CO2EQV. Separate spatially varying aerosol forcings have not yet been implemented. As mentioned in Sect. 2, this trend is treated differently from the other predictors in the sense it is always

included in the empirical model; other predictors are considered only in cases where they appear to add value (following step one of the predictor selection process). Figure 1 shows seasonal correlation between SAT and CO2EQV in the top left panels. Subsequent panels show the correlation derived from predictor–predictand pairs (following removal of the linear trend associated with CO2EQV). Correlation between SAT and CO2EQV is in general strongly positive across the majority of the globe and particularly so when the response of SAT to the internal variability of the climate system is known to be small compared to the response to the signal associated with anthropogenic forcing, for example in the Northern Hemisphere during spring (MAM) and summer (JJA). Additionally, correlation between SAT and
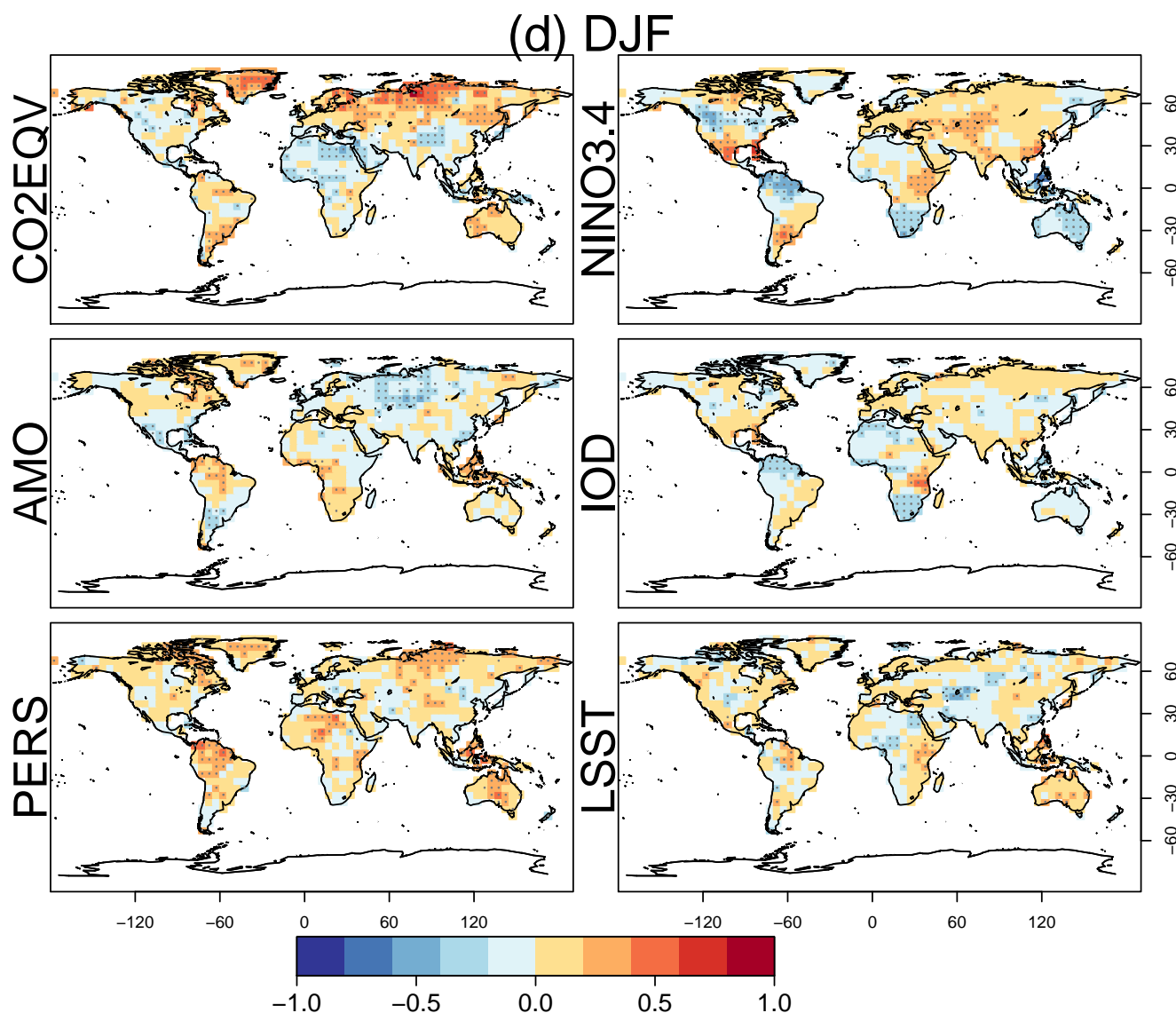
Figure 2. Correlation between PREC and the set of predictors (1961–2013) for **(a)** MAM, **(b)** JJA, **(c)** SON, and **(d)** DJF. As in Fig. 1, correlation with CO2EQV is shown in the top left panel; subsequent panels show correlation between predictand–predictor pairs following removal of the CO2EQV trend. Stippling is used to indicate significance at the 95 % level.

CO2EQV is in general strongly positive throughout tropical land masses at all times of year.

Among the indices describing variability in the climate system, NINO3.4 shows the second strongest relationship with SAT; the importance of ENSO in governing variability in temperatures across the tropics is highlighted by correlations stronger than ±0.5 in parts of South America, Africa and northern Australia in addition to the tropical Pacific and Indian oceans. ENSO-based relationships in extra-tropical land regions are less apparent, although positive correlation in the northern half of the North American continent and negative ones around the Gulf of Mexico show the well-known influence on winter (DJF) and spring (MAM) SAT

(Ropelewski and Halpert, 1987; Kiladis and Diaz, 1989). Very low correlations are found across Europe.

The PDO and IOD correlation patterns are very similar to those for NINO3.4. Much of the signal associated with PDO is likely captured by NINO3.4. However, inclusion of PDO alongside NINO3.4 in the prediction system may yield additional skill in the northern Pacific as a result of enhanced cyclonic circulation around the deepened Aleutian Low associated with a positive, warm PDO phase (Liu and Alexander, 2007). The AMO correlation patterns clearly act independently of ENSO and feature correlations throughout the high northern latitudes and the North Atlantic but curiously not so much in western Europe (van Oldenborgh et al., 2009).
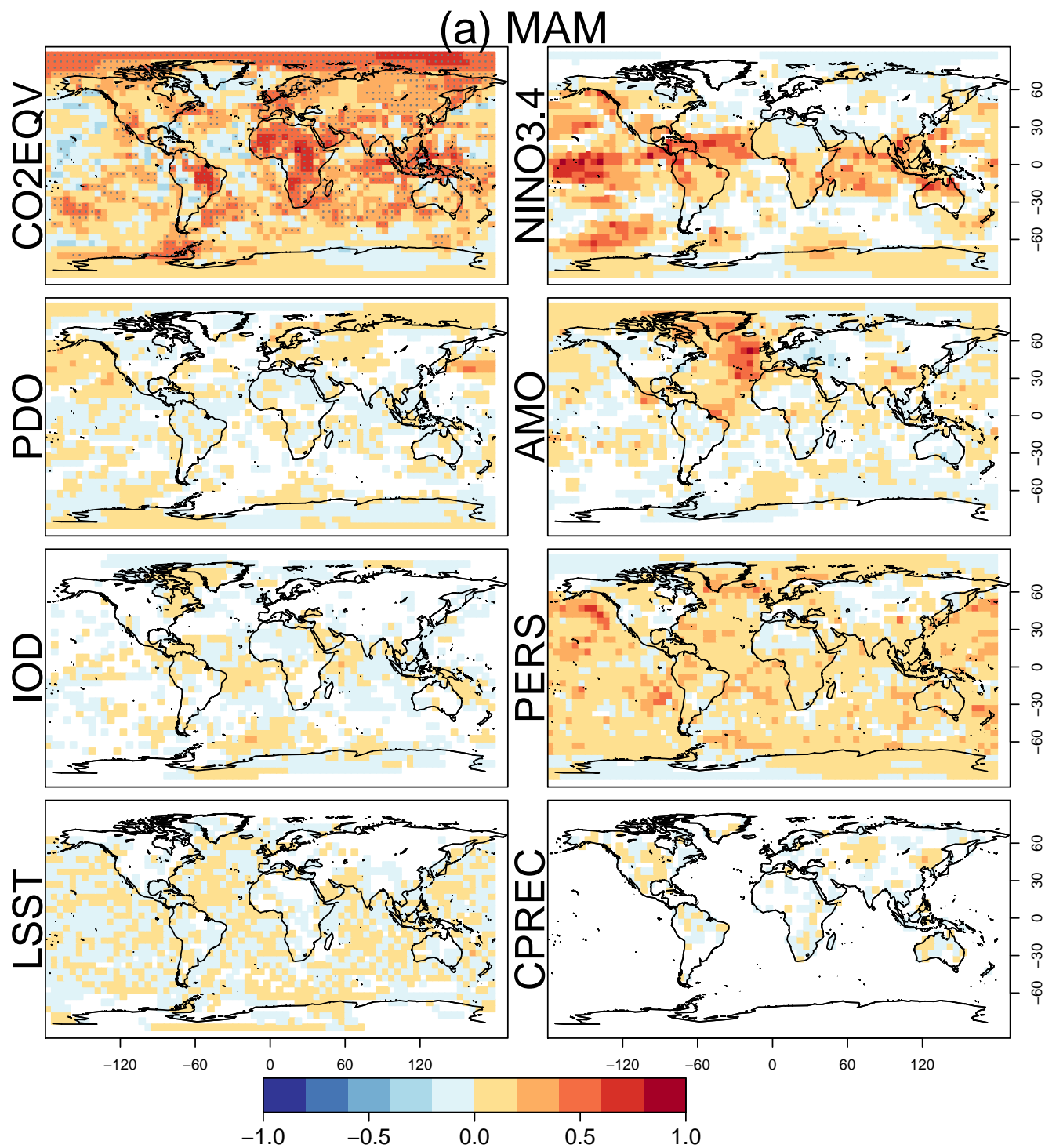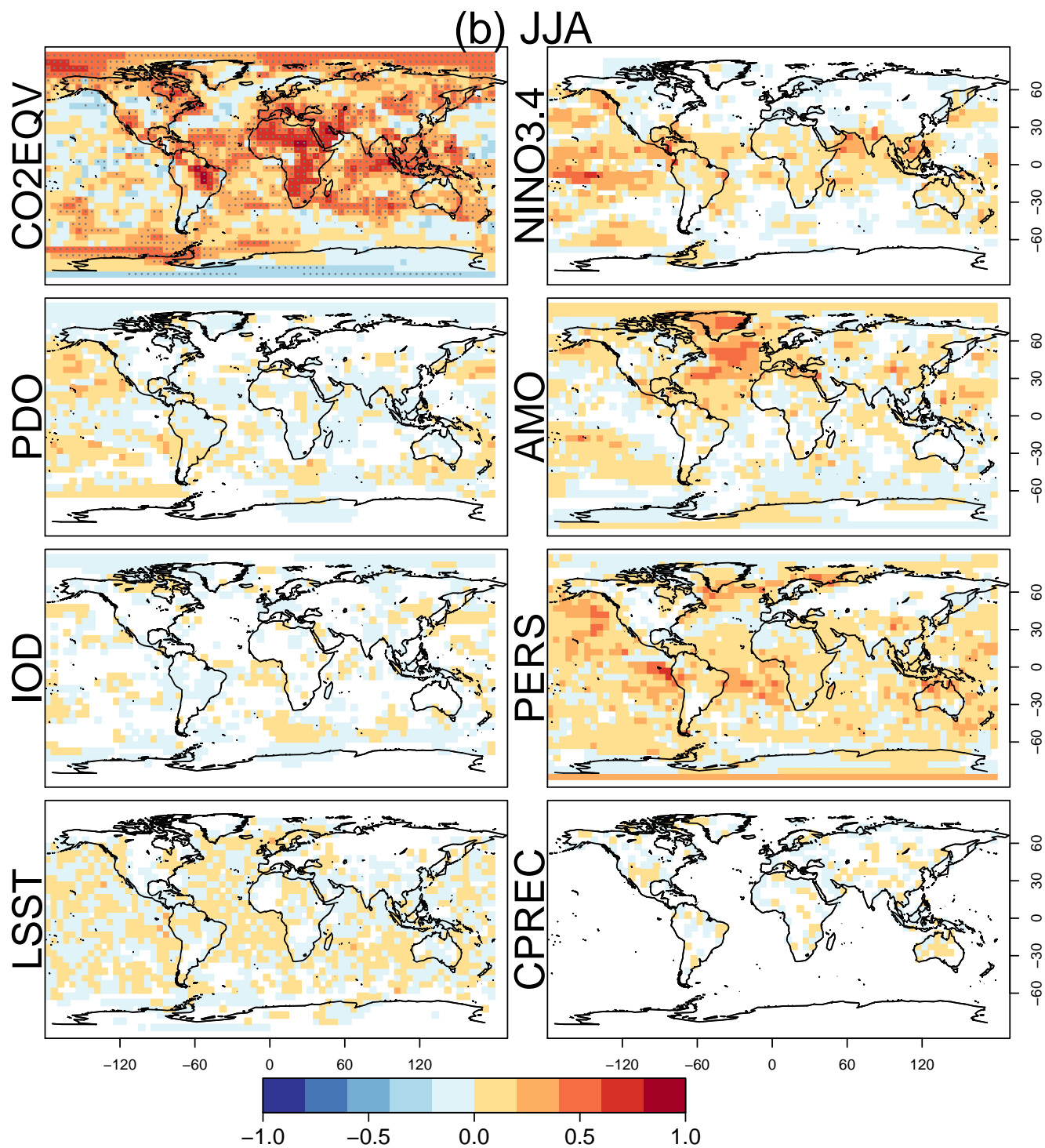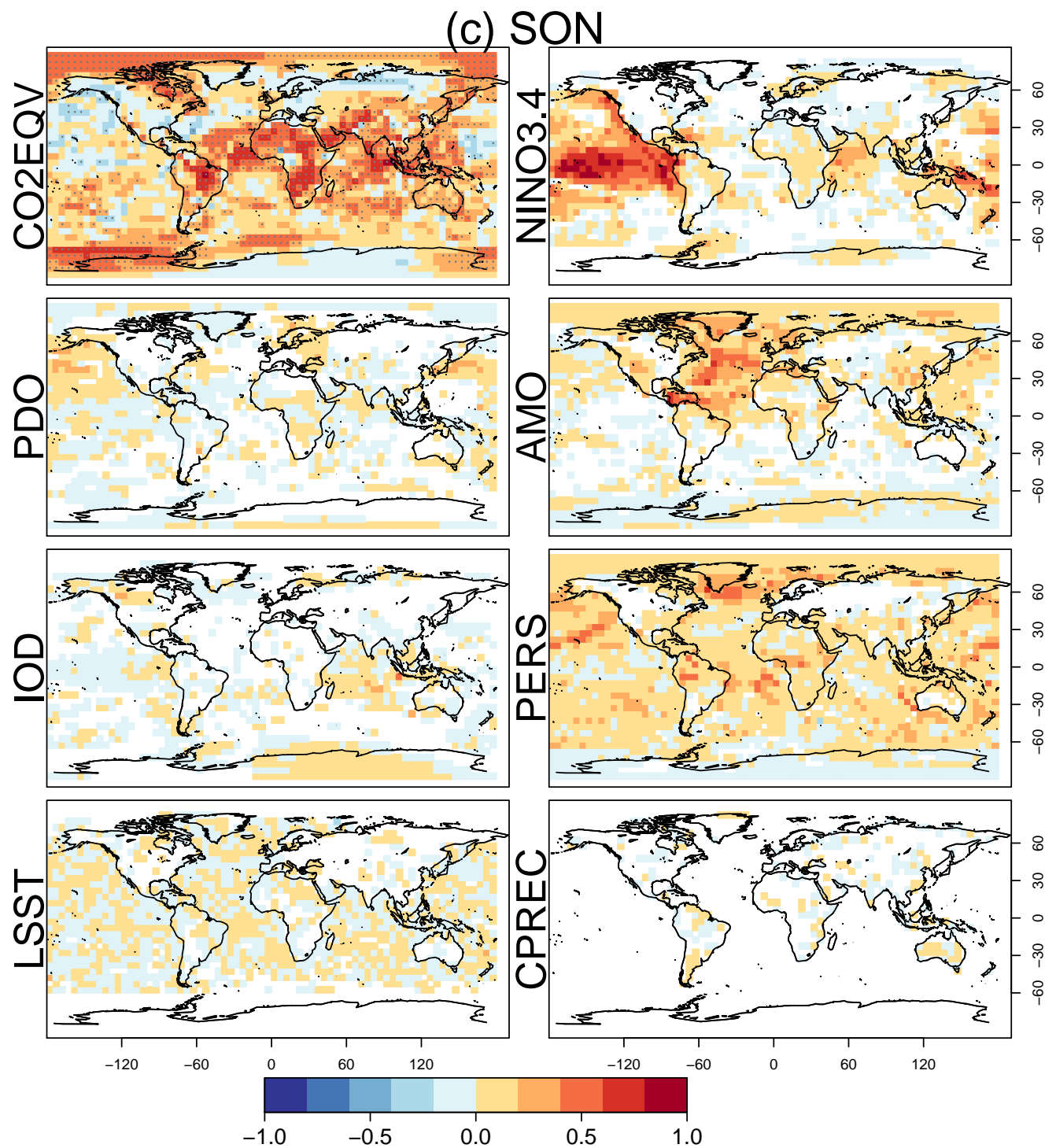
Figure 3.

Figure 3.

Figure 3.
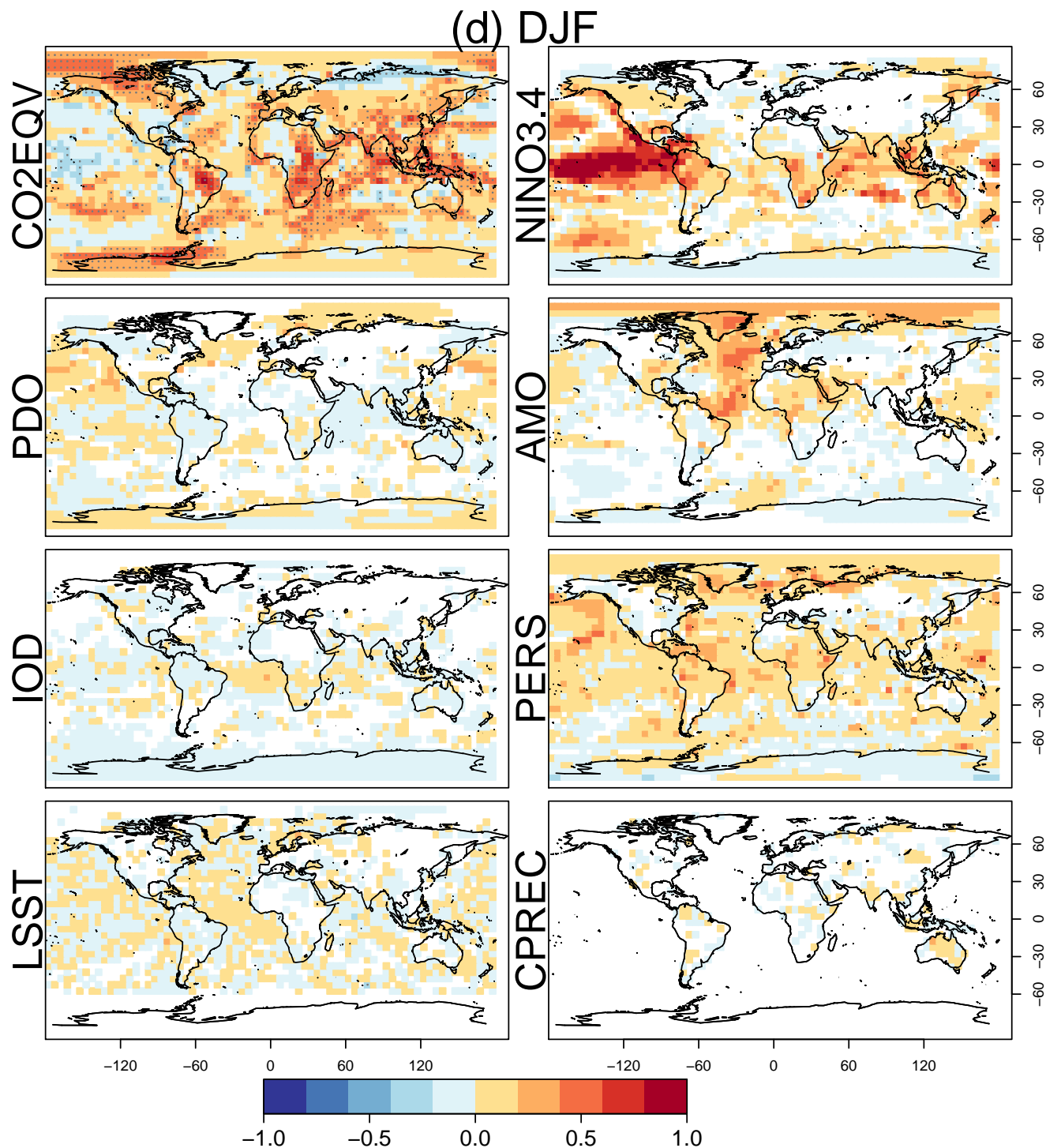
Figure 3. Correlation between SAT hindcasts and observations (1961–2013) for **(a)** MAM, **(b)** JJA, **(c)** SON, and **(d)** DJF. The top left panel shows correlation between observations and SAT hindcasts constructed using the CO_2 equivalent as the sole predictor; stippling is used to indicate significance at the 95 % level. Subsequent panels show differences in correlation following the inclusion of additional predictors.
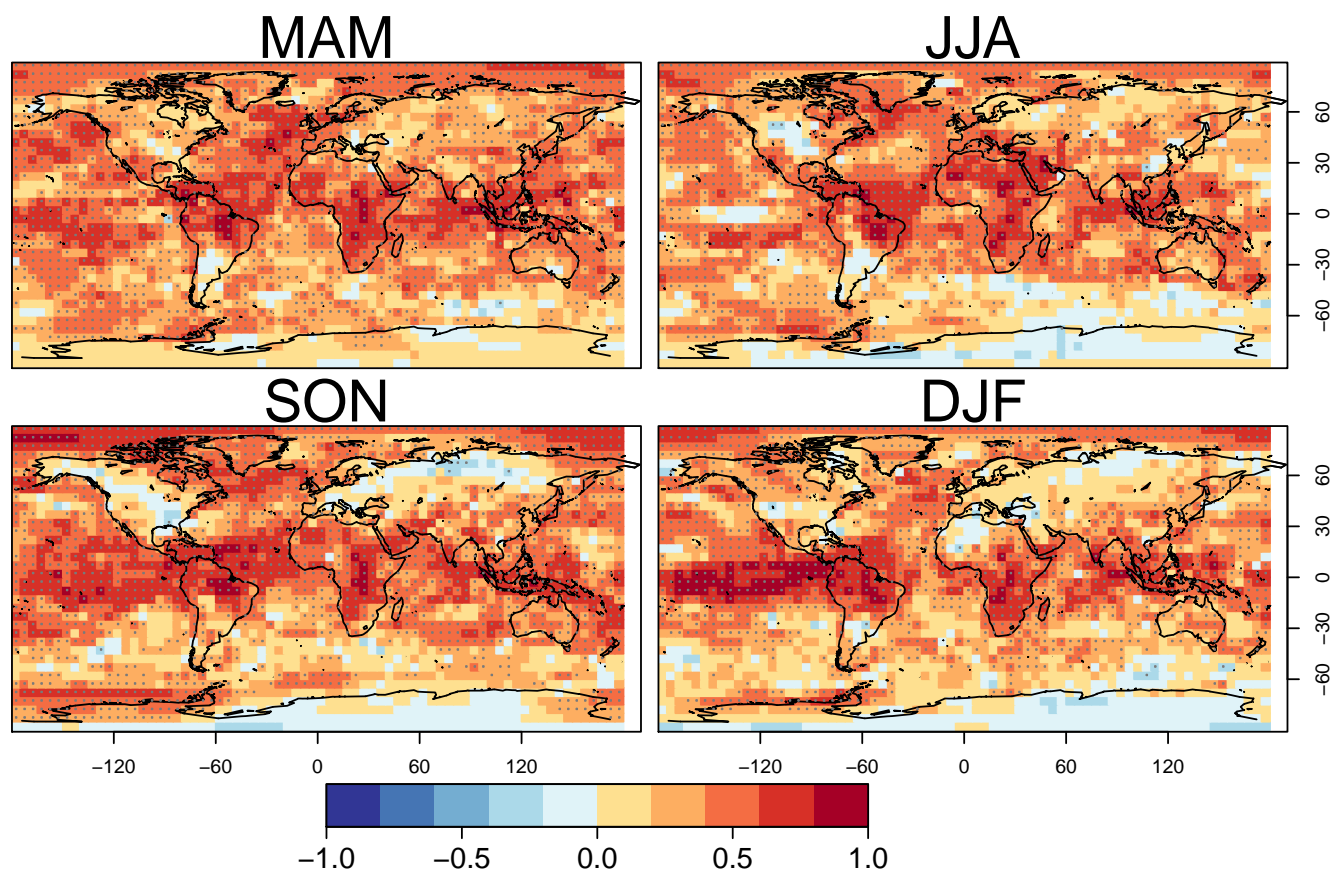
Figure 4. Correlation between observations and SAT hindcasts generated using the regression model with all predictors (1961–2013). Stippling is used to indicate significance at the 95 % level.

The PDO, IOD and AMO indices are all included in the prediction system.

Correlation associated with the QBO is poor with the notable exception of northern and central Russia during the boreal autumn (not shown). In agreement with Folland et al. (2012) we found no significant correlation for winter in Europe with a 1-month lead time. This is surprising given the link found in previous work between the QBO and the Arctic Oscillation (AO), and thus on European surface climate, although the authors suggest that predictability requires a shorter optimal lead time than that used here (Marshall and Scaife, 2009). QBO is thus withdrawn and not included in the prediction system.

Persistence (PERS) shows strong correlations in some key regions and is particularly important for high latitude seas in the Northern Hemisphere during winter, reflecting the latent heat of melting of the sea ice. Over land, however, there are relatively few regions associated with strong correlation outside of the tropics. Correlation is greater than 0.4 in parts of western Europe (MAM), south-east Europe (JJA), central North America (JJA) and parts of central Asia (JJA). However, aside from these examples, the memory of land surface temperature outside of the tropics does not appear to extend to the predictor period.

Local SST (LSST) produces similar correlation to persistence over the oceans but offers no skill over most continental regions. It is anticipated that LSST may be beneficial in coastal regions but this is not clear at the present spatial resolution. Both predictors are made available for selection in the SAT forecast system. The relationship between antecedent precipitation (CPREC) and SAT is in general quite poor but correlation is around 0.3 in northern Europe during spring (MAM), most likely representing the connection of a mild, wet winter to a mild spring. There is negative correlation (although not significant) during summer (JJA) in parts of Europe, which suggests that CPREC is partly able to represent the link between soil moisture and SAT at this time of year shown in previous work (van den Hurk et al., 2012). The correlation is also strong (negative) in parts of Australia and south-east Asia, in addition to southern Africa (MAM) and northern South America (DJF and MAM).
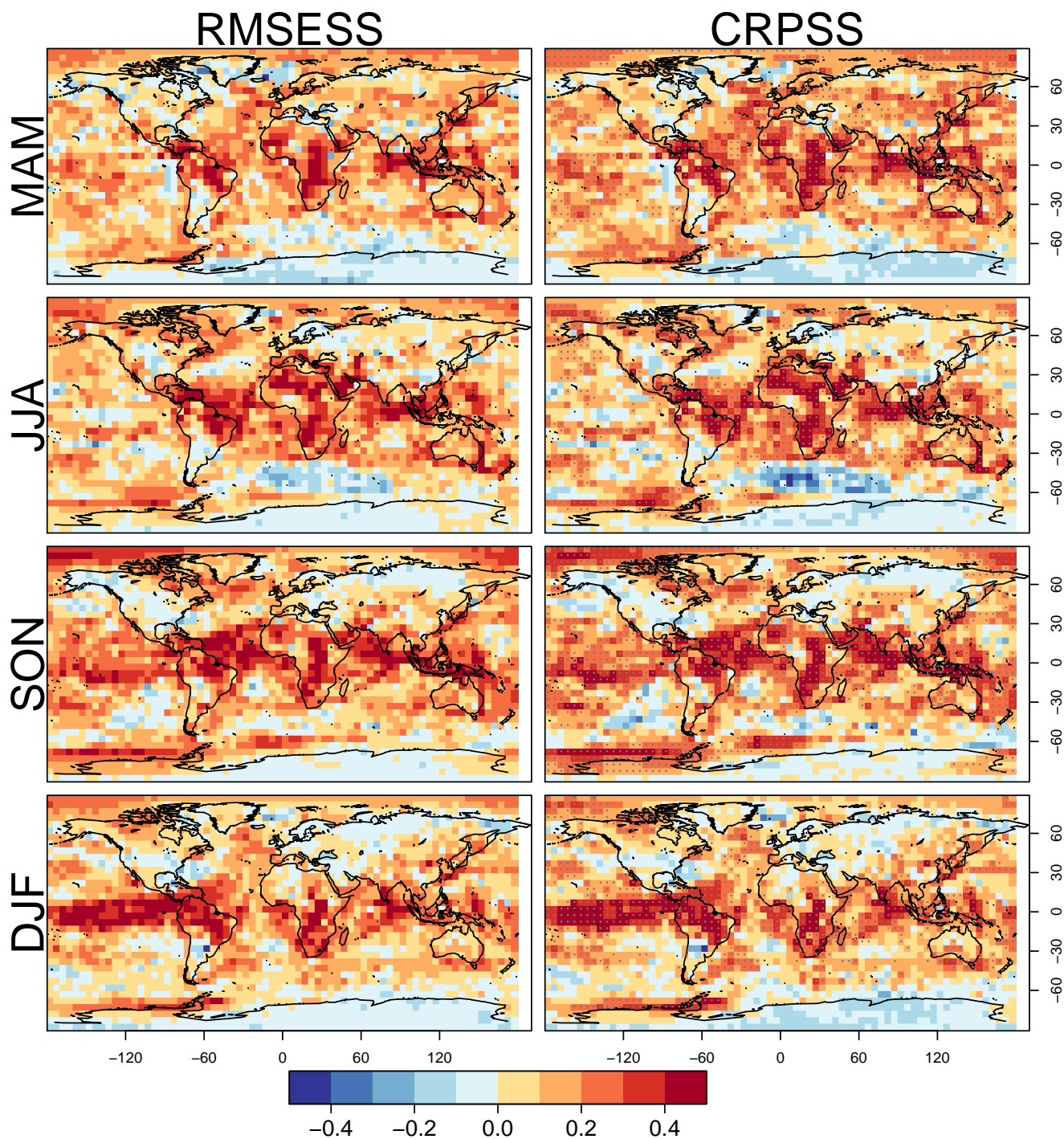
Figure 5. Root-mean-squared error skill score (RMSESS) and the continuous rank probability skill score (CRPSS) of the SAT hindcasts expressed as a skill score against a climatology ensemble forecast (1961–2013). For CRPSS, stippling is used to indicate significance at the 95 % level following a one-sided *t* test.
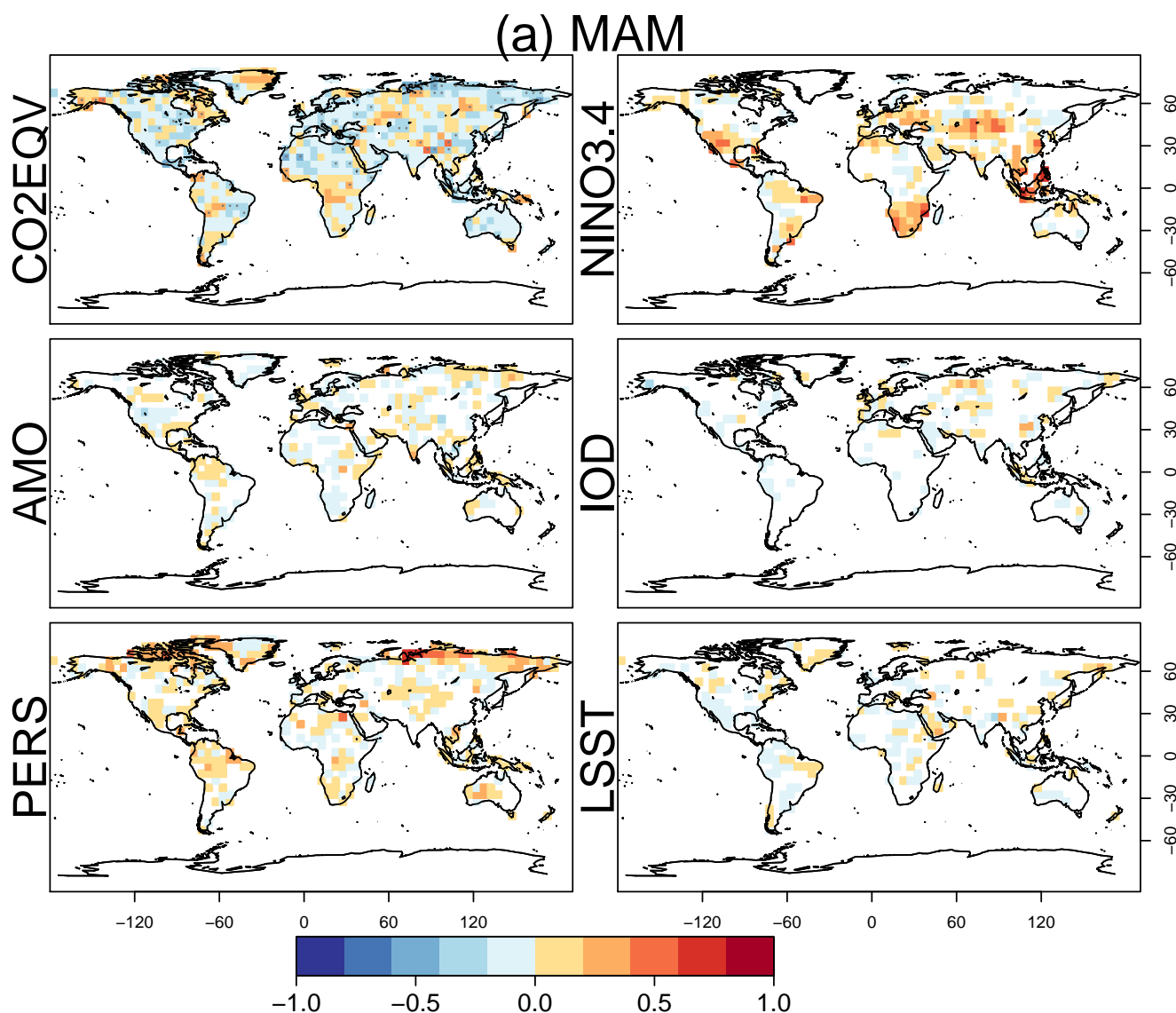
Figure 6.

3.2 Precipitation

Correlation between PREC and the predictors is shown in Fig. 2. As expected, the response of PREC to the trend in CO2EQV is not as strong as that of global temperature. Increased PREC in northern high latitudes during the boreal winter has a known association with global warming (Hartmann et al., 2013). However, the response of precipitation to global warming is not yet visible above the noise in much of the mid-latitudes and these regions are associated with low correlation at all times of the year. Notable exceptions are significant negative correlation in northern Africa (all times of year) and significant positive correlation in Greenland, northern Europe and Asia (MAM, SON and DJF).

The strong correlation exhibited between NINO3.4 and PREC in many parts of the world provides an important basis for predictability. In addition to ENSO-related changes in tropical precipitation patterns, there are a number of known links with precipitation in the extra-tropics (Alexander et al., 2002; Doblas-Reyes et al., 2013), although only a weak one in MAM is found in Europe (van Oldenborgh et al., 2000). Correlation patterns for the PDO (not shown) are again similar for NINO3.4. For the IOD, correlations of around 0.3 exist in eastern Africa during autumn (SON) and winter (DJF), but again these patterns are very similar to those for NINO3.4. Correlation of IOD and PREC following removal of the NINO3.4 signal (not shown) indicates an ENSO-independent relationship, particularly during DJF in eastern Africa, which is supported by the findings of previous work
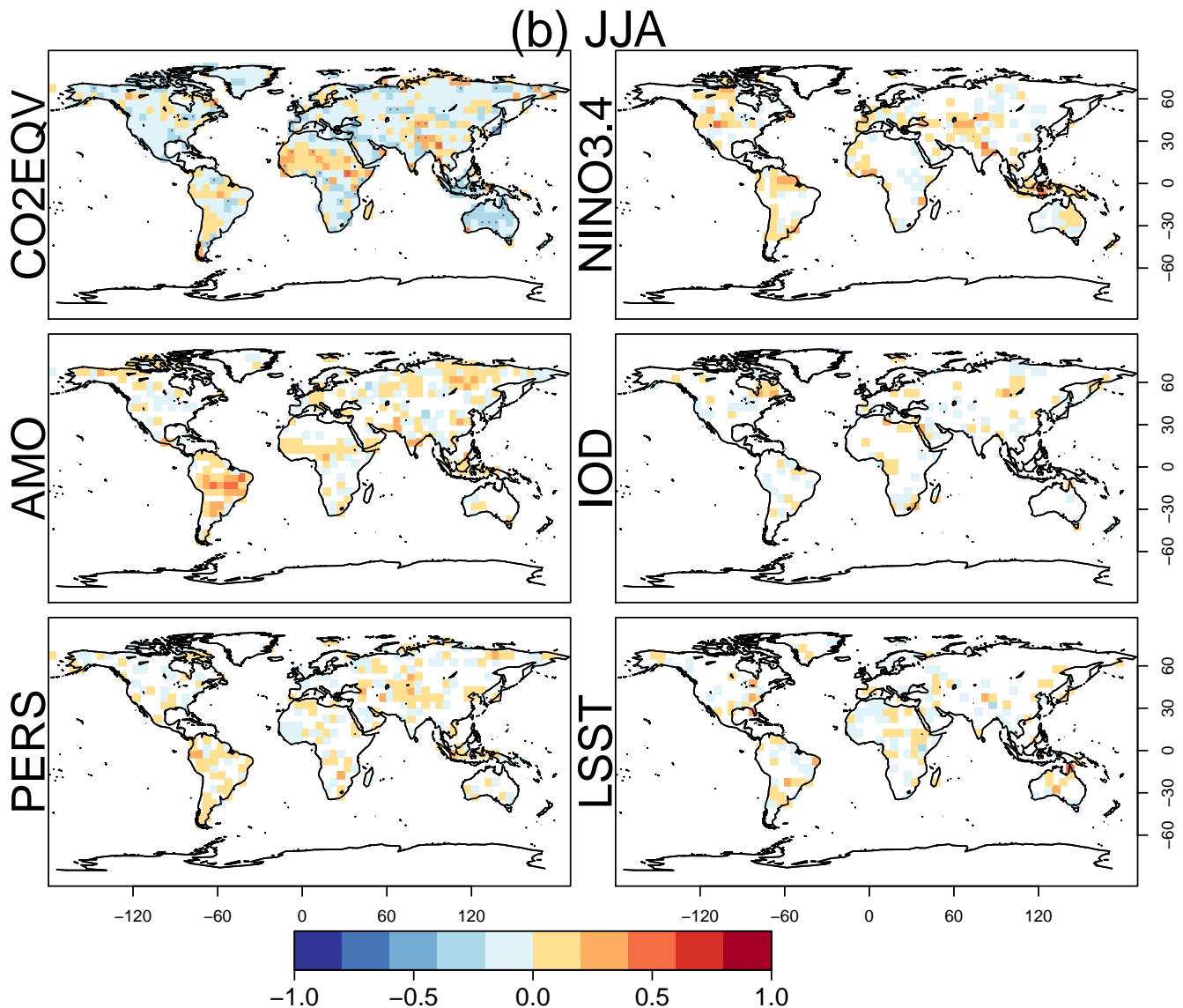
Figure 6.

(Goddard and Graham, 1999), and also parts of Europe. In the absence of known links between the phase of PDO and precipitation anomalies that are independent of ENSO, PDO is not considered for inclusion in the prediction system. QBO is also omitted on the basis that there are few areas of correlation of statistical significance (not shown). AMO on the other hand produces significant correlation in regions influenced by the Atlantic where NINO3.4 does not, including the Sahel (JAS, visible in JJA and SON) and eastern South America (JJA). The AMO–PREC relationship does not appear to extend to extra-tropical regions; there are no discernible areas of strong correlation in Europe or eastern North America. This contrasts with the strong link previously identified between the AMO and JJA precipitation in Europe during the 1990s (Sutton and Dong, 2012). The use of long-term time

series, correlations rather than composites and an absence of temporal filtering here results in lower correlations.

For PERS, there are a number of regions, particularly in the extra-tropics, where significant correlation offers potential for predictability. The most obvious of such correlation is during DJF and MAM in the mid- to high-latitudes of the Northern Hemisphere; the persistence of dry (wet) conditions during autumn in much of central Eurasia is an indicator for similar conditions during winter and into spring. There are relatively few regions where LSST is significantly correlated with PREC. These include the western United States (MAM) and south-east Asia where SST has variability that is independent from ENSO and adds to the skill in dynamical systems (van Oldenborgh et al., 2005a). It remains unclear to
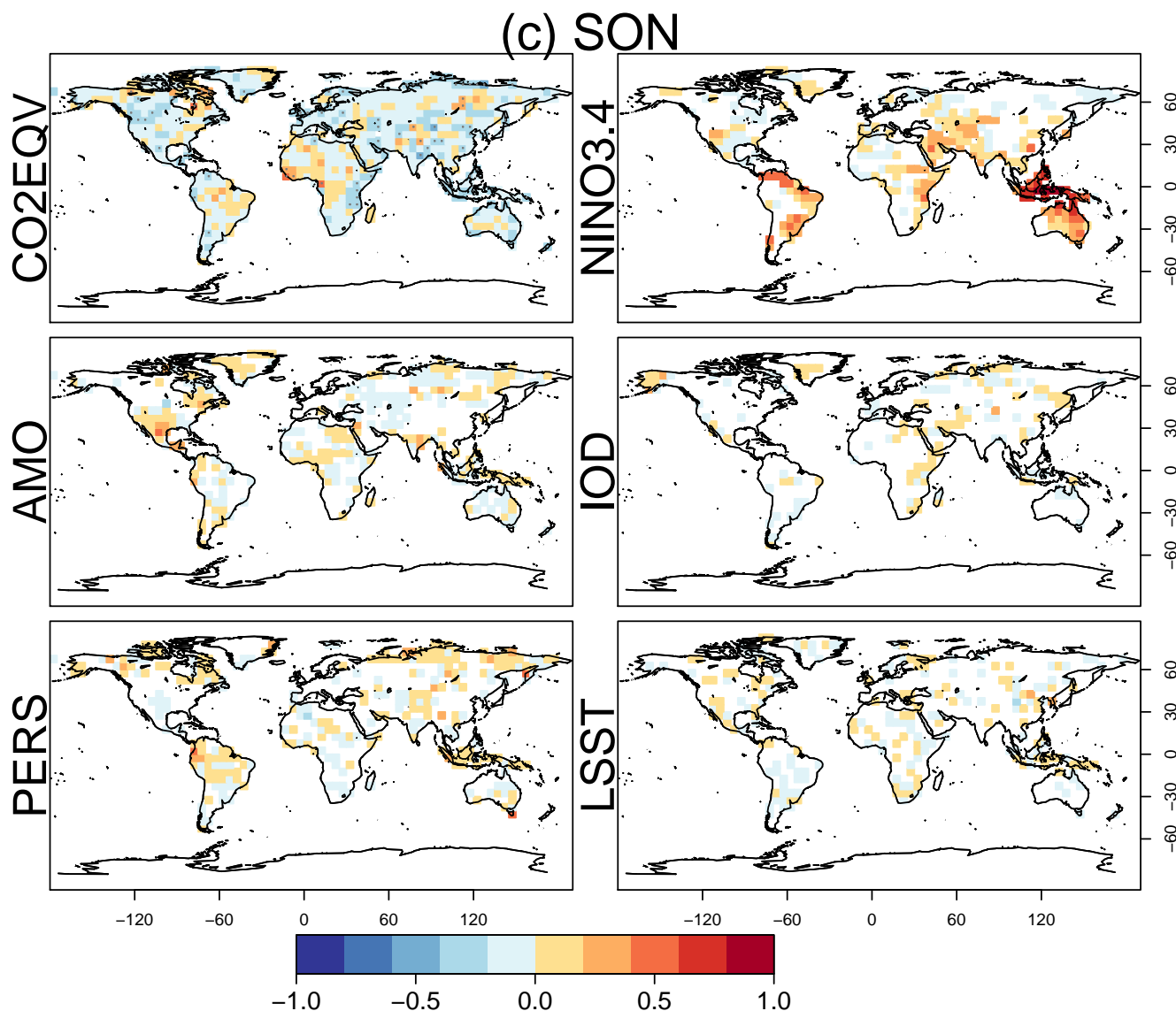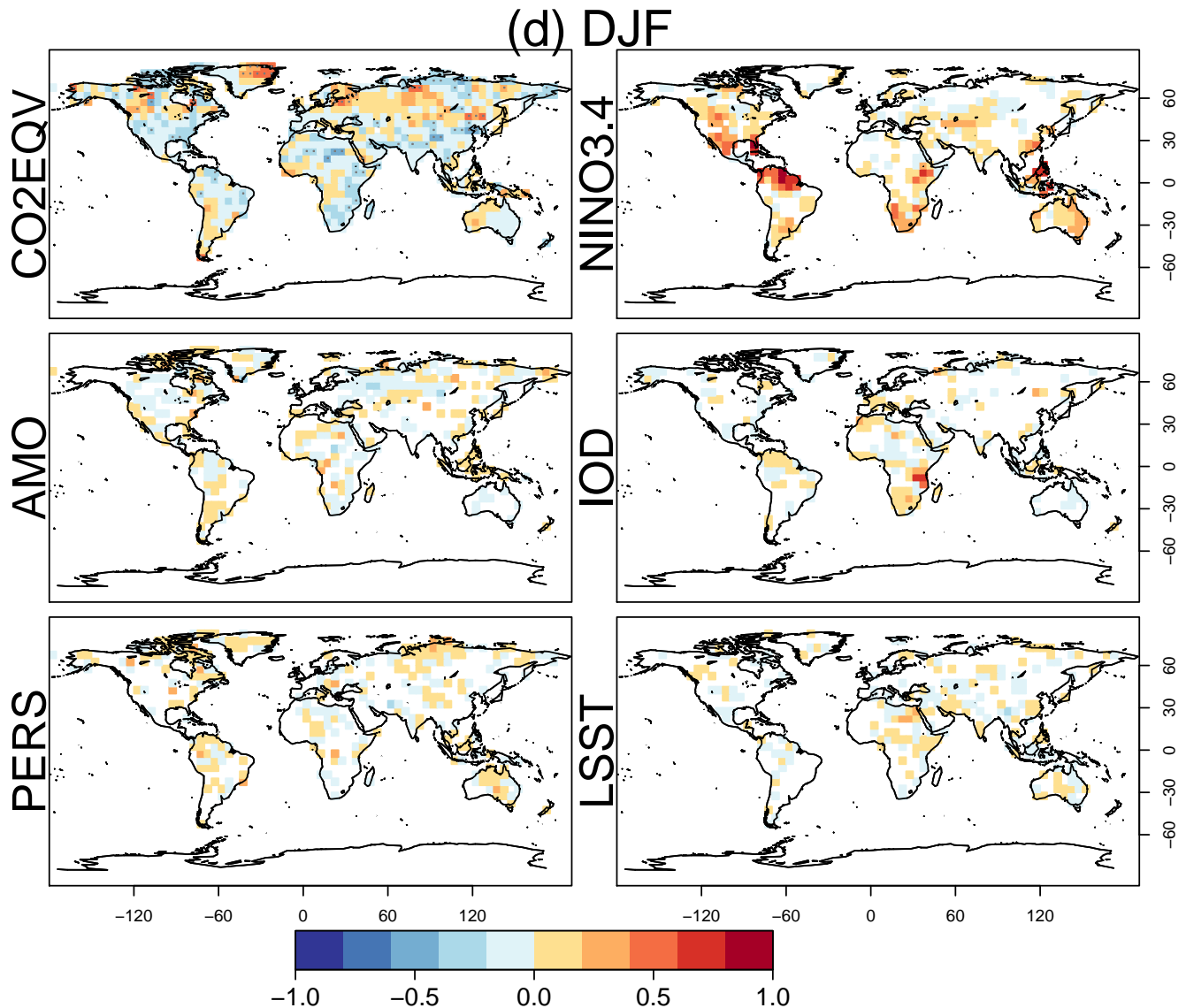
Figure 6.

what extent LSST may offer additional value to this empirical prediction system.

4 Prediction system development and validation

For each hindcast between 1961 and 2013, and for each season and grid point, predictors are selected on the basis of the significance of the (detrended) correlation with the predictand for the fitting period. For validation, causal hindcast estimates are compared with observations to determine the skill of the deterministic and probabilistic aspects of the prediction system.

4.1 Surface air temperature

Following the assessment of potential predictors (step one of the predictor selection process), the following were chosen in addition to CO2EQV for inclusion in the prediction system: NINO3.4, PDO, AMO, IOD, PERS, LSST, and CPREC. Hindcasts were produced with each predictor added in turn and verified against observations. Figure 3 shows the correlation between observations and a hindcast constructed using the CO_2 equivalent only (top left panel on each page) and the incremental correlation attained by including additional predictors cumulatively (subsequent panels on each page). The observation–hindcast correlation following the inclusion of all predictors is given in Fig. 4. Note that these are the correlations of a causal system that only uses information from
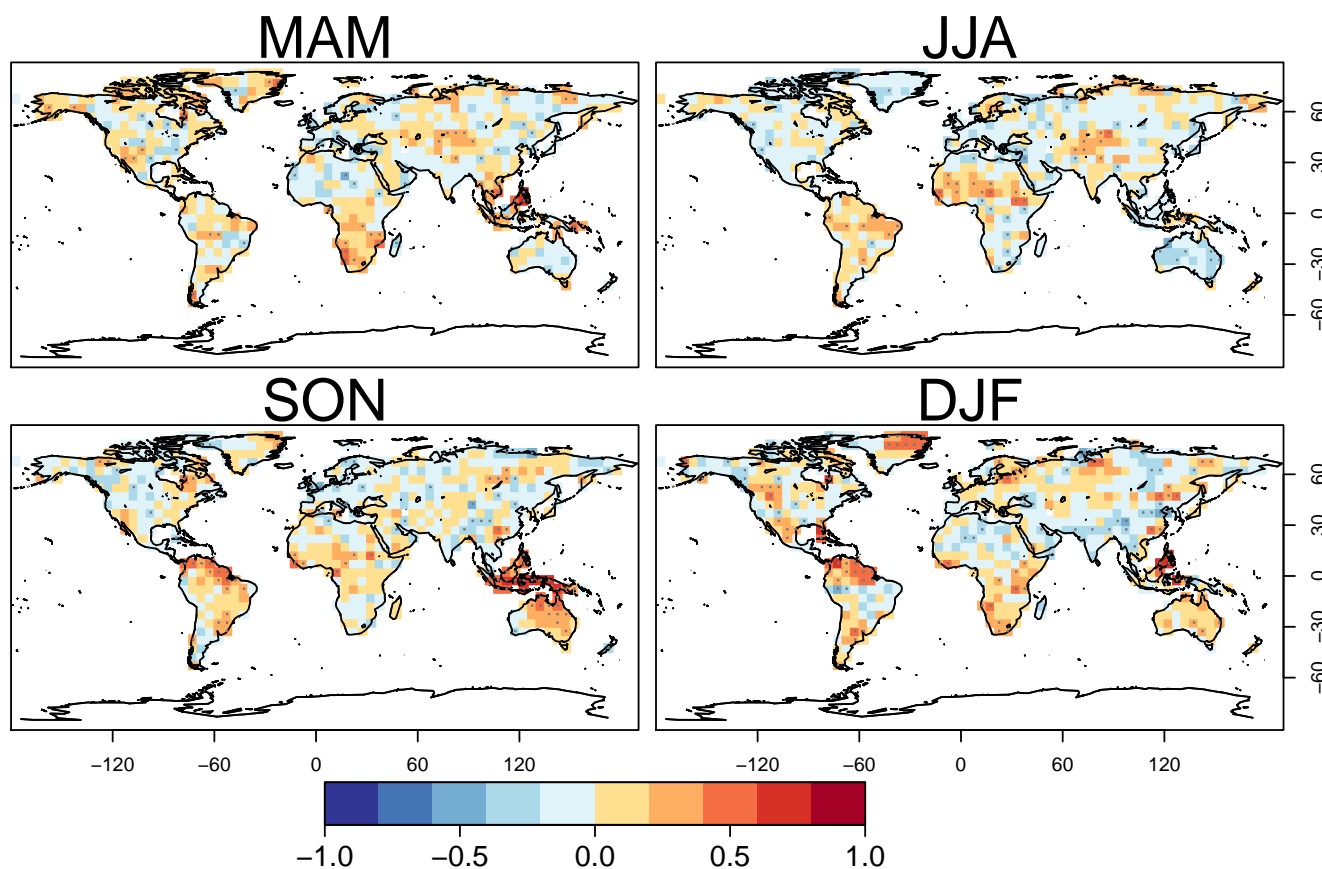
Figure 6. Correlation between PREC hindcasts and observations (1961–2013) for **(a)** MAM, **(b)** JJA, **(c)** SON, and **(d)** DJF. The top left panel shows correlation between observations and SAT hindcasts constructed using the CO_2 equivalent as the sole predictor; stippling is used to indicate significance at the 95 % level. Subsequent panels show difference in correlation following the inclusion of additional predictors.

before the hindcast date, the values are therefore much lower than the full correlations of Fig. 1. If the correlations are spurious – i.e. there was no physical connection but the predictor was included because the correlation exceeded the 90 % significance criterion (this happens by chance on 10 % of the grid points without connection) – the hindcast skill is degraded by the inclusion of this predictor, visible as the light-blue background in the panels of Fig. 3. We tried to minimise this by the first step in the predictor selection process.

The correlation of observations with hindcasts estimated using CO2EQV (Fig. 3) only is much lower than that with hindcasts estimated as a function of all potential predictors (Fig. 4). This is due to the fact that over the first half of the

hindcast period the trend is not yet very strong and does not contribute to the skill. This measure therefore underestimates the skill expected in forecasts, which are made at a time that the trend plays a much larger role, although this depends also on the reference period chosen for the forecasts.

The inclusion of NINO3.4 clearly adds value across the Pacific and in the parts of the tropics. There are no land-based areas where either PDO or IOD add value, but AMO does improve correlation substantially in the North Atlantic and in parts of northern (SON) and eastern (JJA) Europe, although its inclusion degraded the hindcasts in eastern Europe in DJF. The addition of PERS improves correlation in only a handful of locations and LSST; while it is important to corre-

Figure 7. Correlation between observations and PREC hindcasts generated using the regression model with all predictors (1961–2013). Stippling is used to indicate significance at the 95 % level.

lation over some parts of the ocean and hence for islands and coastal regions not resolved by our coarse data sets, it adds little value further from the coast. As suggested in Fig. 1, CPREC adds little global value except in parts of Australia.

The final model shows good skill was found in many regions of the globe (Fig. 4). Key areas of high correlation include the majority of the tropics where the dominance of ENSO on interannual variability is greatest. Correlation is strong at all times of year throughout much of northern South America, central and southern Africa, and south Asia. Strong correlation is also found in important extratropical regions, including much of Europe except during SON. Correlation is strong in much of western and central Europe during the spring and summer (MAM until ASO). Over North America, the skill depends strongly on the season, varying from slightly negative skill (due to overfitting) during SON to good skill in large parts during MAM. Global patterns of RMSE skill scores are broadly similar; regions of strong correlation are generally associated with small differences from observations (Fig. 5; left panels).

Global maps of CRPSS exhibit broad patterns of skill similar to those for correlation (Fig. 5, right panels). The highest skill scores (relative to the climatology-based forecast) are

found in the tropics and are evident during all seasons. In Europe, skill is again greatest during spring and summer, although some parts of eastern Europe and Scandinavia are associated with negative skill scores. Very little of North America is associated with high skill; indeed, the prediction system fails to outperform the climatology-based forecast over the majority of the eastern and southern United States. This lack of skill is known to extend to dynamical forecasts, particularly during winter (e.g. Kim et al., 2012).

4.2 Precipitation

The following predictors were included in the PREC prediction system: NINO3.4, AMO, PERS, and LSST. Figure 6 shows total and incremental correlation results in the same format as Fig. 3 for SAT. Using CO2EQV as a sole predictor fails to yield any notable regions of significant correlation, with the exception of parts of northern Eurasia during winter (DJF). As for SAT, we would expect the forecast skill to be greater than the hindcast skill given that the a large portion of hindcasts were made before the trend becomes important. The addition of NINO3.4 increases hindcast–observation correlation in many parts of the tropics, partic-
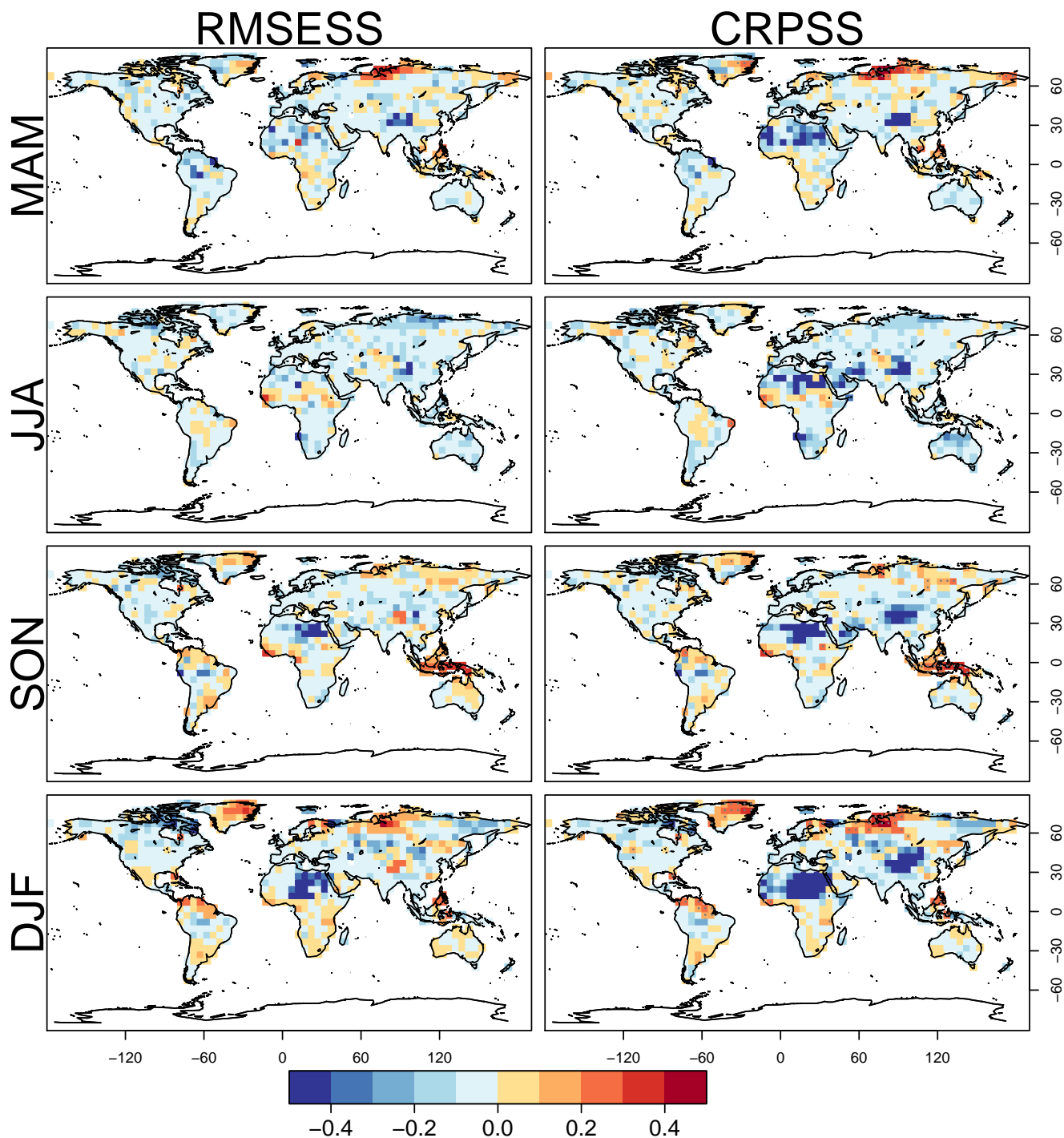
Figure 8. Root-mean-squared error skill score (RMSESS) and the continuous rank probability skill score (CRPSS) of the PREC hindcasts expressed as a skill score against a climatology ensemble forecast (1961–2013). For CRPSS, stippling is used to indicate significance at the 95 % level following a one-sided *t* test.

ularly during the boreal autumn (SON) and winter (DJF). In spite of some evidence for a relationship with PREC in parts of Eurasia as shown in Fig. 2, AMO fails to add any improvement to the empirical model's skill except in northeastern Brazil and to some extent the Sahel. The same is largely true for PERS and LSST, suggesting that almost all skill is captured by NINO3.4 and, to some extent, the climate change signal.

For the final model, high correlation (> 0.6) is limited to south-east Asia and northern parts of South America (between ASO and JFM) (Fig. 7). Another area of high correlation to the north is in south-east South America during the Austral spring (SON–DJF). However, the RMSE for the hindcast is rarely an improvement on that derived from the climatology (Fig. 8, left panels). In addition, there are only a few areas where the hindcast produces a positive CRPSS, which would indicate an improvement on the ensemble forecast derived from the climatology (Fig. 8, right panels). This leads us to conclude that, while the deterministic component of the system is able to reproduce some components of seasonal precipitation variability, probabilistically the system does not perform well outside of limited areas in its present guise.

5 Discussion and outlook

A global empirical system for seasonal climate prediction has been developed and validated. Multiple linear regression was chosen as the basis of the system; a simple predictor selection scheme sought to maximise the predictive skill of a number of predictors describing global-scale modes of variability and local-scale information alongside that of the climate change signal. Probabilistic hindcasts of surface air temperature (SAT) and precipitation (PREC) have been produced using prediction models based on multiple linear regression and validated against observations using correlation and skill scores. The prediction system shows good skill in many regions. For SAT, the trend and interannual variability are well represented throughout the tropics and in a number of extratropical regions, including parts of Europe, particularly during spring and summer, southern Africa and eastern Australia. Skill associated with the probabilistic component of the seasonal predictions shows similar spatial patterns. For PREC, few areas of notable skill are found outside of regions with known ENSO teleconnections and, probabilistically, the system does not perform better than a climatological ensemble throughout most of the world.

As outlined in Sect. 1, the system presented here has been designed to serve both as a benchmark for dynamical prediction systems and as an independent forecast system to be combined with dynamical output to produce more robust forecasts. Concerning the second purpose, it is important to identify seasons and regions where dynamical systems lack skill and to determine whether our system may potentially

add value in such instances. In general, dynamical system skill is limited to regions that are strongly linked to ENSO; in extra-tropical regions, where seasonal variability in the atmospheric state is governed to a greater extent by random internal variability, skill is inevitably lower than in the tropics (Kumar et al., 2007; Arribas et al., 2011). The good skill in many parts of Europe, particularly for forecasts of SAT, is an encouraging property of our system and a detailed comparison with dynamical European forecasts is forthcoming. The inclusion of locally varying predictors, in combination with predictors describing large-scale modes of variability, provides a basis to elicit more skill than can be attained using global indices alone.

An important outcome of this work is the system's implementation in a quasi-operational framework and the provision of regular forecasts. Monthly forecasts are generated for each forthcoming 3-month season and made publicly available through the KNMI (Royal Netherlands Meteorological Institute) Climate Explorer along with uncertainty parameters and updated hindcast validation. The system's framework permits the potential to test empirical prediction methods other than linear regression, such as neural networks that potentially capture non-linear aspects of the climate system. Additionally, as mentioned in Sect. 2, the current list of predictors considered for inclusion is not exhaustive and there is scope to better exploit the predictive information in other locally varying predictors. Further avenues for system development include region-specific and case-based analyses and application to alternative predictands from century-long reanalyses or those describing extreme events. Focus will also be given to alternative methods of ensemble generation using, for instance, derived uncertainty in regression parameters and spatial patterns.

References

Alexander, M. A., Bladé, I., Newman, M., Lanzante, J. R., Lau, N.-C., and Scott, J. D.: The atmospheric bridge: The influence of ENSO teleconnections on air-sea interaction over the global oceans, J. Climate, 15, 2205–2231, 2002.

Arribas, A., Glover, M., Maidens, A., Peterson, K., Gordon, M., MacLachlan, C., Graham, R., Fereday, D., Camp, J., Scaife, A. A., Xavier, P., McLean, P., Colman, A., and Cusack, S. : The GloSea4 ensemble prediction system for seasonal forecasting, Mon. Weather Rev., 139, 1891–1910, 2011.

Baldwin, M. P., Gray, L. J., Dunkerton, T. J., Hamilton, K., Haynes, P. H., Randel, W. J., Holton, J. R., Alexander, M. J., Hirota, I.,

Horinouchi, T., Jones, D. B. A., Kinnersley, J. S., Marquardt, C., Sato, K., and Takahashi, M.: The quasi-biennial oscillation, Rev. Geophys., 39, 179–229, 2001.

Balmaseda, M. and Anderson, D.: Impact of initialization strategies and observations on seasonal forecast skill, Geophys. Res. Lett., 36, L01701, doi:10.1029/2008GL035561, 2009.

Brands, S., Manzanas, R., Gutiérrez, J. M., and Cohen, J.: Seasonal predictability of wintertime precipitation in Europe using the snow advance index, J. Climate, 25, 4023–4028, 2012.

Brönnimann, S., Annis, J. L., Vogler, C., and Jones, P. D.: Reconstructing the quasi-biennial oscillation back to the early 1900s, Geophys. Res. Lett., 34, L22805, doi:10.1029/2007GL031354, 2007.

Chevallier, M. and Salas-Mélia, D.: The role of sea ice thickness distribution in the Arctic sea ice potential predictability: A diagnostic approach with a coupled GCM, J. Climate, 25, 3025–3038, 2012.

Coelho, C. A. S., Stephenson, D. B., Balmaseda, M., Doblas-Reyes, F. J., and van Oldenborgh, G. J.: Toward an integrated seasonal forecasting system for South America, J. Climate, 19, 3704–3721, doi:10.1175/JCLI3801.1, 2006.

Cowtan, K. and Way, R. G.: Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends, Q. J. Roy. Meteor. Soc., 140, 1935–1944, 2014.

Doblas-Reyes, F. J., Hagedorn, R., Palmer, T. N., and Morcrette, J.-J.: Impact of increasing greenhouse gas concentrations in seasonal ensemble forecasts, Geophys. Res. Lett., 33, L07708, doi:10.1029/2005GL025061, 2006.

Doblas-Reyes, F. J., Garcia-Serrano, J., Lienert, F., Pinto Biescas, A., and Rodrigues, L. R. L.: Seasonal climate predictability and forecasting: status and prospects, Wiley Interdisciplinary Reviews – Climate Change, 4, 245–268, doi:10.1002/wcc.217, 2013.

Dutra, E., Schär, C., Viterbo, P., and Miranda, P.: Land-atmosphere coupling associated with snow cover, Geophys. Res. Lett., 38, L15707, doi:10.1029/2011GL048435, 2011.

Ebdon, R. A. and Veryard, R. G.: Fluctuations in Equatorial Stratospheric Winds, Nature, 189, 791–793, 1961.

Ferro, C. A. T.: Fair scores for ensemble forecasts, Q. J. Roy. Meteor. Soc., 140, 1917–1923, doi:10.1002/qj.2270, 2013.

Folland, C. K., Scaife, A. A., Lindesay, J., and Stephenson, D. B.: How potentially predictable is northern European winter climate a season ahead?, Int. J. Climatol., 32, 801–818, 2012.

Goddard, L. and Graham, N. E.: Importance of the Indian Ocean for simulating rainfall anomalies over eastern and southern Africa, J. Geophys. Res.-Atmos. (1984–2012), 104, 19099–19116, 1999.

Hartmann, D. L., Klein Tank, A. M. G., Rusticucci, M., Alexander, L. V., Brönnimann, S., Charabi, Y., Dentener, F. J., Dlugokencky, E. J., Easterling, D. R., Kaplan, A., Soden, B. J., Thorne, P. W., Wild, M., and Zhai, P. M.: Observations: Atmosphere and Surface, in: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, A., and Midgley, P. M., Cambridge University Press, Cambridge, UK and New York, NY, USA, 159–254, doi:10.1017/CBO9781107415324.008, 2013.

Ho, C. K., Hawkins, E., Shaffrey, L., and Underwood, F. M.: Statistical decadal predictions for sea surface temperatures: a bench-mark for dynamical GCM predictions, Clim. Dynam., 41, 917–935, doi:10.1007/s00382-012-1531-9, 2013.

Kennedy, J. J., Rayner, N. A., Smith, R. O., Parker, D. E., and Saunby, M.: Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 1. Measurement and sampling uncertainties, J. Geophys. Res.-Atmos. (1984–2012), D14103, doi:10.1029/2010JD015218, 116, 2011a.

Kennedy, J. J., Rayner, N. A., Smith, R. O., Parker, D. E., and Saunby, M.: Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. Biases and homogenization, J. Geophys. Res.-Atmos. (1984–2012), 116, D14104, doi:10.1029/2010JD015220, 2011b.

Kiladis, G. N. and Diaz, H. F.: Global climatic anomalies associated with extremes in the Southern Oscillation, J. Climate, 2, 1069–1090, 1989.

Kim, H.-M., Webster, P. J., and Curry, J. A.: Seasonal prediction skill of ECMWF System 4 and NCEP CFSv2 retrospective forecast for the Northern Hemisphere Winter, Clim. Dynam., 39, 2957–2973, 2012.

Krueger, O. and Von Storch, J.-S.: A Simple Empirical Model for Decadal Climate Prediction, J. Climate, 24, 1276–1283, 2011.

Kumar, A., Jha, B., Zhang, Q., and Bounoua, L.: A new methodology for estimating the unpredictable component of seasonal atmospheric variability, J. Climate, 20, 3888–3901, 2007.

Liniger, M. A., Mathis, H., Appenzeller, C., and Doblas-Reyes, F. J.: Realistic greenhouse gas forcing and seasonal forecasts, Geophys. Res. Lett., 34, L04705, doi:10.1029/2006GL028335 , 2007.

Liu, Z. and Alexander, M.: Atmospheric bridge, oceanic tunnel, and global climatic teleconnections, Rev. Geophys., 45, RG2005, doi:10.1029/2005RG000172, 2007.

Marshall, A. G. and Scaife, A. A.: Impact of the QBO on surface winter climate, J. Geophys. Res.-Atmos. (1984–2012), 114, D18110, doi:10.1029/2009JD011737, 2009.

Mason, S. J. and Mimmack, G. M.: Comparison of some statistical methods of probabilistic forecasting of ENSO, J. Climate, 15, 8–29, 2002.

Meinshausen, M., Smith, S. J., Calvin, K., Daniel, J. S., Kainuma, M. L. T., Lamarque, J.-F., Matsumoto, K., Montzka, S. A., Raper, S. C. B., Riahi, K., Thomson, A., Velders, G. J. M., and van Vuuren, D. P. P.: The RCP greenhouse gas concentrations and their extensions from 1765 to 2300, Climatic Change, 109, 213–241, 2011.

Morice, C. P., Kennedy, J. J., Rayner, N. A., and Jones, P. D.: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set, J. Geophys. Res.-Atmos. (1984–2012), 117, D08101, doi:10.1029/2011JD017187, 2012.

Newman, M.: An Empirical Benchmark for Decadal Forecasts of Global Surface Temperature Anomalies, J. Climate, 26, 5260–5269, 2013.

Peng, P., Kumar, A., Barnston, A. G., and Goddard, L.: Simulation skills of the SST-forced global climate variability of the NCEP-MRF9 and the Scripps-MPI ECHAM3 models, J. Climate, 13, 3657–3679, 2000.

Penland, C. and Matrosova, L.: Prediction of tropical Atlantic sea surface temperatures using linear inverse modeling, J. Climate, 11, 483–496, 1998.

Quesada, B., Vautard, R., Yiou, P., Hirschi, M., and Seneviratne, S. I.: Asymmetric European summer heat predictability from wet and dry southern winters and springs, Nature Climate Change, 2, 736–741, 2012.

Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., Kent, E. C., and Kaplan, A.: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century, J. Geophys. Res.-Atmos. (1984–2012), 108, 4407, doi:10.1029/2002JD002670, 2003.

Ropelewski, C. F. and Halpert, M. S.: Global and Regional Scale Precipitation Patterns Associated with the El Niño/Southern Oscillation, Mon. Weather Rev., 115, 1606–1626, 1987.

Sardeshmukh, P. D., Compo, G. P., and Penland, C.: Changes of probability associated with El Nino, J. Climate, 13, 4268–4286, 2000.

Schepen, A., Wang, Q. J., and Robertson, D. E.: Combining the strengths of statistical and dynamical modeling approaches for forecasting Australian seasonal rainfall, J. Geophys. Res.-Atmos. (1984–2012), 117, D20107, doi:10.1029/2012JD018011, 2012.

Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A., Rudolf, B., and Ziese, M.: GPCC Full Data Reanalysis Version 6.0 at 2.5°: Monthly Land-Surface Precipitation from Rain-Gauges built on GTS-based and Historic Data, Global Precipitation Climatology Centre (GPCC) at Deutscher Wetterdienst, doi:10.5676/DWD_GPCC/FD_M_V6_250, available at: http://gpcc.dwd.de/ (last access: January 2015), 2011.

Shongwe, M. E., Ferro, C. A. T., Coelho, C. A. S., and van Oldenborgh, G. J.: Predictability of cold spring seasons in Europe, Mon. Weather Rev., 135, 4185–4201, 2007.

Sterl, A., van Oldenborgh, G. J., Hazeleger, W., and Burgers, G.: On the robustness of ENSO teleconnections, Clim. Dynam., 29, 469–485, 2007.

Suckling, E. B. and Smith, L. A.: An Evaluation of Decadal Probability Forecasts from State-of-the-Art Climate Models, J. Climate, 26, 9334–9347, 2013.

Sutton, R. T. and Dong, B.: Atlantic Ocean influence on a shift in European climate in the 1990s, Nat. Geosci., 5, 788–792, 2012.

van den Dool, H.: Empirical methods in short-term climate prediction, Oxford University Press, New York, NY, USA, 2007.

van den Hurk, B., Doblas-Reyes, F., Balsamo, G., Koster, R. D., Seneviratne, S. I., and Camargo Jr, H.: Soil moisture effects on seasonal temperature and precipitation forecast scores in Europe, Clim. Dynam., 38, 349–362, 2012.

van Oldenborgh, G. J., Burgers, G., and Klein Tank, A.: On the El Niño teleconnection to spring precipitation in Europe, Int. J. Climatol., 20, 565–574, doi:10.1002/(SICI)1097-0088(200004)20:5<565::AID-JOC488>3.0.CO;2-5, 2000.

van Oldenborgh, G. J., Balmaseda, M. A., Ferranti, L., Stockdale, T. N., and Anderson, D. L. T.: Evaluation of atmospheric fields from the ECMWF seasonal forecasts over a 15-year period, J. Climate, 18, 3250–3269, 2005a.

van Oldenborgh, G. J., Balmaseda, M. A., Ferranti, L., Stockdale, T. N., and Anderson, D. L. T.: Did the ECMWF seasonal forecast model outperform statistical ENSO forecast models over the last 15 years?, J. Climate, 18, 3240–3249, 2005b.

van Oldenborgh, G. J., te Raa, L. A., Dijkstra, H. A., and Philip, S. Y.: Frequency- or amplitude-dependent effects of the Atlantic meridional overturning on the tropical Pacific Ocean, Ocean Sci., 5, 293–301, doi:10.5194/os-5-293-2009, 2009.

Weisheimer, A. and Palmer, T. N.: On the reliability of seasonal climate forecasts, J. R. Soc. Interface, 11, 20131162, doi:10.1098/rsif.2013.1162, 2014.

Weisheimer, A., Doblas-Reyes, F. J., Palmer, T. N., Alessandri, A., Arribas, A., Déqué, M., Keenlyside, N., MacVean, M., Navarra, A., and Rogel, P.: ENSEMBLES: A new multi-model ensemble for seasonal-to-annual predictions – Skill and progress beyond DEMETER in forecasting tropical Pacific SSTs, Geophys. Res. Lett., 36, L21711, doi:10.1029/2009GL040896, 2009.

Zhao, M. and Hendon, H. H.: Representation and prediction of the Indian Ocean dipole in the POAMA seasonal forecast model, Q. J. Roy. Meteor. Soc., 135, 337–352, 2009.