Karlsruhe Institute of Technology

Assessing Predictive Performance: From Precipitation Forecasts over the Tropics to Receiver Operating Characteristic Curves and Back

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

von der KIT-Fakultät für Mathematik des
Karlsruher Instituts für Technologie (KIT)
genehmigte

DISSERTATION

von

M.Sc. Peter Vogel

aus

Marburg

Tag der mündlichen Prüfung: 06.02.2019

Referent: Prof. Dr. Tilmann Gneiting
1. Korreferent: Prof. Dr. Peter Knippertz
2. Korreferent: Prof. Dr. Daniel S. Wilks

Abstract

Educated decision making involves two major ingredients: probabilistic forecasts for future events or quantities and an assessment of predictive performance. This thesis focuses on the latter topic and illustrates its importance and implications from both theoretical and applied perspectives.

Receiver operating characteristic (ROC) curves are key tools for the assessment of predictions for binary events. Despite their popularity and ubiquitous use, the mathematical understanding of ROC curves is still incomplete. We establish the equivalence between ROC curves and cumulative distribution functions (CDFs) on the unit interval and elucidate the crucial role of concavity in interpreting and modeling ROC curves. Under this essential requirement, the classical binormal ROC model is strongly inhibited in its flexibility and we propose the novel beta ROC model as an alternative. For a class of models that includes the binormal and the beta model, we derive the large sample distribution of the minimum distance estimator. This allows for uncertainty quantification and statistical tests of goodness-of-fit or equal predictive ability. Turning to empirical examples, we analyze the suitability of both models and find empirical evidence for the increased flexibility of the beta model. A freely available software package called `betaROC` is currently prepared for release for the statistical programming language `R`.

Throughout the tropics, probabilistic forecasts for accumulated precipitation are of economic importance. However, it is largely unknown how skillful current numerical weather prediction (NWP) models are at timescales of one to a few days. For the first time, we systematically assess the quality of nine global operational NWP ensembles for three regions in northern tropical Africa, and verify against station and satellite-based observations and for the monsoon seasons 2007–2014. All examined NWP models are uncalibrated and unreliable, in particular for high probabilities of precipitation, and underperform in the prediction of amount and occurrence of precipitation when compared to a climatological reference forecast. Statistical postprocessing corrects systematic deficiencies and realizes the full potential of ensemble forecasts. Postprocessed forecasts are calibrated and reliable and outperform raw ensemble forecasts in all regions and monsoon seasons. Disappointingly however, they have predictive performance only equal to the climatological reference. This assessment is robust and holds for all examined NWP models, all monsoon seasons, accumulation periods of 1 to 5 days, and station and spatially aggregated satellite-based observations. Arguably, it implies that current NWP ensembles cannot translate information about the atmospheric state into useful information regarding occurrence or amount of precipitation. We suspect convective parameterization as likely cause of the poor performance of NWP ensemble forecasts as it has been shown to be a first-order error source for

the realistic representation of organized convection in NWP models.

One may ask if the poor performance of NWP ensembles is exclusively confined to northern tropical Africa or if it applies to the tropics in general. In a comprehensive study, we assess the quality of two major NWP ensemble prediction systems (EPSs) for 1 to 5-day accumulated precipitation for ten climatic regions in the tropics and the period 2009–2017. In particular, we investigate their skill regarding the occurrence and amount of precipitation as well as the occurrence of extreme events. Both ensembles exhibit clear calibration problems and are unreliable and overconfident. Nevertheless, they are (slightly) skillful for most climates when compared to the climatological reference, except tropical and northern arid Africa and alpine climates. Statistical postprocessing corrects for the lack of calibration and reliability, and improves forecast quality. Postprocessed ensemble forecasts are skillful for most regions except the above mentioned ones.

The lack of NWP forecast skill in tropical and northern arid Africa and alpine climates calls for alternative approaches for the prediction of precipitation. In a pilot study for northern tropical Africa, we investigate whether it is possible to construct skillful statistical models that rely on information about recent rainfall events. We focus on the prediction of the probability of precipitation and find clear evidence for its modulation by recent precipitation events. The spatio-temporal correlation of rainfall coincides with meteorological assumptions, is reasonably pronounced and stable, and allows to construct meaningful statistical forecasts. We construct logistic regression based forecasts that are reliable, have a higher resolution than the climatological reference forecast, and yield an average improvement of 20% for northern tropical Africa and the period 1998–2014.

Acknowledgements

For the great guidance and support that I received over the last three and a half years, I want to express my deep gratitude to Tilmann Gneiting, Peter Knippertz, and Andreas H. Fink. They truly live the spirit of interdisciplinary research and provided an environment that allowed me to pursue challenging theoretical and applied statistical questions. I have greatly enjoyed my doctoral studies under their excellent supervision and I am thankful for the many small and large things I learned from them. Additionally, I want to express my gratitude to Daniel S. Wilks for agreeing to be a reviewer of this Ph.D. thesis.

The research presented in this thesis has been funded by the German Science Foundation through the Transregional Collaborative Research Center SFB/TRR 165 "Waves to Weather". Furthermore, I am glad about the infrastructural support that I received from the Karlsruhe Institute of Technology and the Heidelberg Institute of Theoretical Studies.

Over the course of my doctoral studies, I have benefited from the knowledge and experience of many colleagues. I would like to thank Andreas Schlueter, Sebastian Lerch, Florian Pantillon, Linda Schneider, Constanze Wellmann, Philipp Zschenderlein, and all other early career scientists in "Waves to Weather" for many fruitful discussions, Gregor Pante, Marlon Maranan, and all members of the working group Atmospheric Dynamics for sharing their expertise in meteorology with me, and Werner Ehm, Alexander Jordan, Stefan Hemri, Kira Feldmann, Fabian Krüger, Patrick Schmidt, Roman Schefzik, Michael Scheuerer and Manuel Klar for numerous stimulating mathematical discussions and insights. Furthermore, I want to thank Ken Mylne, Tom Hamill, Martin Leutbecher, Peter Bechthold, Zied Bouallègue, and Roberto Buizza for excellent feedback and discussions. I always appreciated the friendly and welcoming atmosphere at the Institute of Meteorology and Climate Research and at the Institute for Stochastics of the Karlsruhe Institute of Technology and at the Heidelberg Institute of Theoretical Studies, and feel glad to have such open minded and friendly colleagues.

Finally, I want to thank Anna Carolina, my family, and my friends. I feel very fortunate in light of their enduring support.

List of abbreviations

AE	Absolute error
AEW	African easterly wave
AMMA	African Monsoon Multidisciplinary Analysis
AUC	Area under the ROC curve
BMA	Bayesian model averaging
BS	Brier score
CDF	Cumulative distribution function
CEP	Conditional event probability
CMA	China Meteorological Administration
CPTEC	Centro de Previsão Tempo e Estudos Climáticos
CRPS	Continuous ranked probability score
CNT	Control (run)
ECMWF	European Centre for Medium-Range Weather Forecasts
ENS	Perturbed ensemble (runs)
EMOS	Ensemble model output statistics
EPC	Extended probabilistic climatology
EPS	Ensemble prediction system
FAR	False alarm rate
GEV	Generalized extreme value (distribution)
GTS	Global telecommunication system
HIV	Human immunodeficiency virus
HR	Hit rate
HRES	High-resolution (run)
KASS-D	Karlsruhe African surface station database
KMA	Korea Meteorological Administration
KS	Kolmogorov–Smirnov (distance)
LR	Likelihood ratio
MAE	Mean absolute error
MCS	Mesoscale convective system
MD(E)	Minimum distance (estimation/estimator)
MF	Météo France
MSC	Meteorological Service of Canada
NCEP	National Centres for Environmental Prediction
NWP	Numerical weather prediction
SVM	Support vector machine
TIGGE	The International Grand Global Ensemble
PAV	Pool-adjacent-violators
PIT	Probability integral transform

PoP	Probability of precipitation
RMM	Reduced multi-model (ensemble)
ROC	Receiver operating characteristic
SEEPS	Stable equitable error in probability space
TRMM	Tropical Rainfall Measuring Mission
UKMO	UK Met Office
uPIT	Unified probability integral transform
UTC	Universal time coordinated
WMO	World Meteorological Organization

Contents

1 Introduction

At all times, it has been a desire of mankind to learn more about the inherently uncertain future and to obtain an idea of how it might look like and change our lives. Predictions for future events or quantities can thereby act as guidance and facilitate sound decisions. Educated decision making that takes uncertainties into account necessarily requires predictions that are probabilistic in nature. Scientifically supported by fundamental results on the chaotic nature of many processes (e.g., Lorenz, 1963), forecasting has experienced a transition from a deterministic to a probabilistic approach. Nowadays, probabilistic forecasting is state-of-the-art in various fields of application including, but not limited to, meteorology, hydrology, economics, and demography.

With the introduction of probabilistic forecasts arose the need for theoretically principled tools for their verification. Chapter 2 introduces fundamental concepts and key tools for the assessment of probabilistic forecasts. Of particular interest in many situations is the assessment of predictions for binary outcomes. ROC curves are key tools in these settings, but despite their popularity and ubiquitous use, the mathematical understanding of ROC curves is still incomplete. Chapter 3 introduces ROC curves and their properties and advances the mathematical understanding. In particular, we argue that the class of ROC curves and the class of CDFs on the unit interval are equivalent and elucidate the essential role of concavity for the interpretation of ROC curves. Moving from the theoretical analysis of ROC curves to their modeling, we analyze shortcomings of the classical binormal and related ROC curve models and propose the novel beta ROC model as alternative. For parameter estimation, we rely on minimum distance (MD) estimation and derive the asymptotic distribution of the MD estimator for a class of models that includes the classical binormal and the novel beta ROC model. This allows then for uncertainty quantification and statistical tests of goodness-of-fit or equal predictive ability. On empirical examples, we analyze the suitability of the binormal and the beta ROC model and propose extensions for the latter to account for specific features of ROC curves that are commonly found in practice.

Chapters 5 and 6 focus on the assessment of probabilistic forecasts for precipitation. To this end, Chapter 4 introduces briefly the concepts of NWP models, ensembles and EPSs, and statistical postprocessing. To evaluate the quality of NWP ensemble forecasts, we further construct a probabilistic climatology and investigate its properties. In a nutshell, NWP models describe atmospheric processes by partial differential equations and are the state-of-the-art approach to predict future weather. Started from slightly different initial conditions, ensembles are a set of deterministic NWP model forecasts where each ensemble member

represents a potential realization of the future state of the atmosphere. Despite the advance in the formulation of NWP models and the setup of EPSs, systematic errors remain and require statistical postprocessing to reveal the full potential of ensemble forecasts.

One particular challenge in weather forecasting is the prediction of accumulated precipitation, especially when it is related to moist convection. Chapter 5 investigates the quality of raw and postprocessed ensemble forecasts from nine global NWP models for accumulated precipitation in three regions in northern tropical Africa. To obtain a complete assessment, we verify predictions for 1–5 day accumulated precipitation against station and satellite observations at various spatial aggregations for the period 2007–2014. All results reveal clear deficiencies of raw ensemble forecasts for accumulated precipitation, clear improvements in ensemble forecast quality by statistical postprocessing, but even after postprocessing hardly any skill when compared to a climatological reference forecast.

Based on this comprehensive assessment of NWP ensemble forecast skill for accumulated precipitation in northern tropical Africa, one can ask if the poor performance of NWP ensembles is exclusively confined to this region or if it applies to the tropics in general. In Chapter 6 we evaluate the quality of raw and postprocessed forecasts from two global NWP ensembles for amount and occurrence of precipitation as well as the occurrence of extreme rainfall. We partition the tropical land mass based on climatic properties and verify forecasts against satellite-based observations that allow for a consistent assessment of forecast quality throughout the tropics. For accumulation periods of 1–5 days, raw ensemble forecasts suffer from the same deficiencies as in northern tropical Africa. Nevertheless, they are slightly skillful in many regions for the period 2009–2017 relative to a probabilistic climatology, and statistical postprocessing further improves forecast skill. From 2009 to 2017, the improvement in NWP forecast quality is mostly small and even postprocessed forecasts for precipitation do not outperform the climatological reference for tropical and northern arid Africa and in complex terrain.

In Chapter 7, we briefly investigate alternative approaches for forecasting the occurrence of precipitation in northern tropical Africa. We analyze the spatio-temporal correlation of precipitation and detect clear modulations of the probability of precipitation by recent rainfall events. Subsequently, logistic regression based forecasts for the prediction of precipitation are constructed. The evaluation across 1998–2014 reveals clear improvements of our approach relative to climatological and NWP ensemble forecasts.

Chapter 8 summarizes and discusses key results of this dissertation and provides an outlook to future research.

1.1 Relation to previous and published work

As suggested by the title “Assessing Predictive Performance: From Precipitation Forecasts over the Tropics to Receiver Operating Characteristic Curves and

Back", the order in which research results are presented in this thesis differs from the chronological one. The following list of research articles is in chronological order and all research presented in this thesis contains significant contributions by myself.

Vogel et al. (2018) Vogel, P., Knippertz, P., Fink, A. H., Schlueter, A. and Gneiting, T. (2018). Skill of global raw and postprocessed ensemble predictions of rainfall over northern tropical Africa. *Weather and Forecasting*, **33**, 369–388.

This research articles forms the basis of Chapter 5 and contributes to parts of Chapters 2 and 4. Its copyright belongs to the American Meteorological Society (AMS).[1]

Gneiting and Vogel (2018) Gneiting, T. and Vogel, P. (2018). Receiver operating characteristic (ROC) curves. Preprint, `arXiv:1809.04808`.

Chapter 3 is almost identical to this research article. It contains additional results on notions of forecast dominance as well as a generalization of an impossibility result for the concavity of ROC curves.

The work presented in Chapters 6 and 7 is based on joint, ongoing research with Tilmann Gneiting, Peter Knippertz, Andreas H. Fink, and Andreas Schlueter.

[1]

2 Preliminaries on forecasting and verification

For many decades, predictions have been deterministic in the form of point forecasts. While conveying information about future events, they lack information about the uncertainty inherent to the prediction. Motivated by the fundamental results of Lorenz (1963) and others on the chaotic and non-linear nature underlying many key applications of forecasting, a shift in paradigms towards probabilistic forecasting has occurred (e.g., Gneiting, 2008; Gneiting and Katzfuss, 2014). In meteorology, NWP ensembles have been introduced in the 1990s (e.g., Toth and Kalnay, 1993; Buizza et al., 2000) and have become state-of-the-art for generating probabilistic forecasts (see Chapter 4).

The rise of probabilistic forecasts necessitated the study of theoretically principled tools for their evaluation. In this chapter, we review fundamentals of probabilistic forecast assessment, in particular the concepts of prediction spaces, calibration, reliability, proper scoring rules, and consistent scoring functions.

2.1 Prediction spaces

Murphy and Winkler (1987) introduced a mathematical framework for the evaluation of point forecasts based on the joint distribution of observations and forecasts, that Gneiting and Ranjan (2013) extended to accommodate probabilistic forecasts. We follow Gneiting and Ranjan (2013) and consider the joint distribution of multiple probabilistic forecasts and an observation on a probability space $(\Omega, \mathcal{A}, \mathbb{Q})$. We assume that elements of the sample space Ω can be identified by tuples

$$(P_1, \ldots, P_k, Y),$$

where each of $P_1, \ldots, P_k$ is a probability measure on the outcome space $(\Omega_Y, \mathcal{A}_Y)$ of the observation Y. Further let each P_i, $i = 1, \ldots, k$, be measurable with respect to the sub-σ-algebra $\mathcal{A}_i \subseteq \mathcal{A}$ that encodes the information a forecast is based on. We restrict the discussion to the case of real-valued observations, so that $(\Omega_Y, \mathcal{A}_Y) = (\mathbb{R}, \mathcal{B})$, and identify each P_i with its associated right-continuous CDF F_i.

In this particular setting, the elements of Ω can be identified by tuples

$$(F_1, \ldots, F_k, Y, V),$$

where Y is a real-valued random variable. V has a standard uniform distribution, is independent of $\mathcal{A}_1, \ldots, \mathcal{A}_k$ and Y, and allows to assess the calibration of CDF-valued forecasts with discontinuities and ensemble forecasts in (2.3) and (2.4). In the following, we consider often only one probabilistic forecast which we then denote by F.

2.2 Calibration, sharpness, reliability, and resolution

Probabilistic forecasts are meant to provide information about future events. As such, they should convey correct probabilistic statements, in that observations behave like random draws from the forecast distributions. This property is called *probabilistic calibration* and while other notions of calibration exist (Gneiting et al., 2007), it is considered the most critical requirement for CDF-valued probabilistic forecasts (Dawid, 1984; Diebold et al., 1998; Gneiting et al., 2007). Under all probabilistically calibrated forecasts, sharper forecasts with lesser uncertainty are preferred.

For a CDF-valued continuous random quantity F, the *probability integral transform* (PIT) is

$$Z_F = F(Y) \tag{2.1}$$

and probabilistic calibration of F is defined via a standard uniform distribution of the PIT Z_F. In applications, probabilistic calibration is assessed empirically via PIT histograms. For a test set

$$\{(F_i, y_i) \,|\, i = 1, \ldots, n\}$$

representing a sample of size n from the joint distribution of the forecast and observation, a PIT histogram displays the PIT values of all n forecast–observation pairs. For probabilistically calibrated forecasts, the PIT histogram is uniform, and for miscalibrated forecasts, information on the type of miscalibration is displayed in the PIT histogram. Commonly encountered in applications are underdispersed forecasts that have too little variance. Consequently, the observations fall too frequently into the tails of the forecast distribution and the PIT histogram has a U-shape. For overdispersed forecasts, one observes hump-shaped histograms, while skewness of the PIT histogram indicates a bias. In applications, one often observes both dispersion errors and biases as exemplarily displayed in Figure 2.1.

Frequently, probabilistic forecasts for real-valued quantities are given by a discrete sample $f_{ij}, j = 1, \ldots, m$, where each f_{ij} is drawn from F_i, $j = 1, \ldots, m$. Prominent examples are NWP ensemble forecasts where each of the m ensemble members is generated in a slightly different fashion and represents a potential realization of Y. For historic reasons, the calibration of such forecasts is assessed via *verification rank* or *Talagrand histograms* (Anderson, 1996; Talagrand et al., 1997; Hamill and Colucci, 1997). The verification rank

$$r_i = \#\{j \,|\, f_{ij} < y_i\} + 1$$

is the rank of the observation when it is pooled with the m ensemble members. If the forecast that generated the simple random sample $f_{i1}, \ldots, f_{im}$ is probabilistically calibrated, then r_i is uniformly distributed on the set of possible ranks $\{1, \ldots, m+1\}$ with

$$\text{prob}(r_i = j) = \frac{1}{m+1} \quad \text{for } j = 1, \ldots, m+1. \tag{2.2}$$

Deviations from a uniform distribution indicate miscalibration with the same interpretation as for PIT histograms.

In many applications such as for temperature or pressure, the distribution of the observation Y is continuous. For precipitation, however, there is a positive probability of no precipitation and the definition of PIT and verification rank histograms needs suitable adaptions. For the PIT, the definition is extended to encompass discontinuities of F in that

$$Z_F = F(Y-) + V(F(Y) - F(Y-)), \tag{2.3}$$

assigns a random value between the right-hand and the left-hand limit $F(y-) = \lim_{x \uparrow y} F(x)$ of F at any point of discontinuity y. With this extension, the equivalence is recovered as proven by Rüschendorf (2009).

In case of an ensemble forecast with $k > 1$ ensemble members predicting the value of the verifying observation, define a minimum tied rank $r_{i,\min}$ and a maximum tied rank $r_{i,\max}$ by

$$r_{i,\min} := \#\{j \mid f_{ij} < y_i\} + 1 \quad \text{and} \quad r_{i,\max} := \#\{j \mid f_{ij} \leq y\} + 1.$$

The verification rank r_i is then a random draw from $\{r_{i,\min}, \ldots, r_{i,\max}\}$. In case of precipitation, if k ensemble member predict no precipitation and no precipitation is observed, then r_i is a random draw between 1 and $k+1$.

In Chapters 5 and 6, we compare discrete probabilistic forecasts from different ensembles against each other as well as postprocessed forecasts in form of CDFs. With varying numbers of ensemble members and hence bins in the verification rank histograms, a visual comparison between different ensemble forecasts as well as between ensemble and postprocessed forecasts is difficult.

To allow a compelling visual assessment of calibration in this setting, we use *unified probability integral transform* (uPIT) histograms as introduced by Vogel et al. (2018). For a CDF-valued forecast F, the uPIT Z'_F is simply the PIT Z_F. For an ensemble forecast with m members, compute the verification rank r_i and define the uPIT Z'_F as

$$Z'_F := \frac{r_i - 1}{m+1} + \frac{V}{m+1}, \tag{2.4}$$

where V is standard uniform and independent of the forecast and observation.

Figure 2.1 displays in the top row verification rank histograms for ensemble forecasts from the China Meteorological Administration (CMA) with 14 members, the UK Met Office (UKMO) with 23 members, and the European Centre

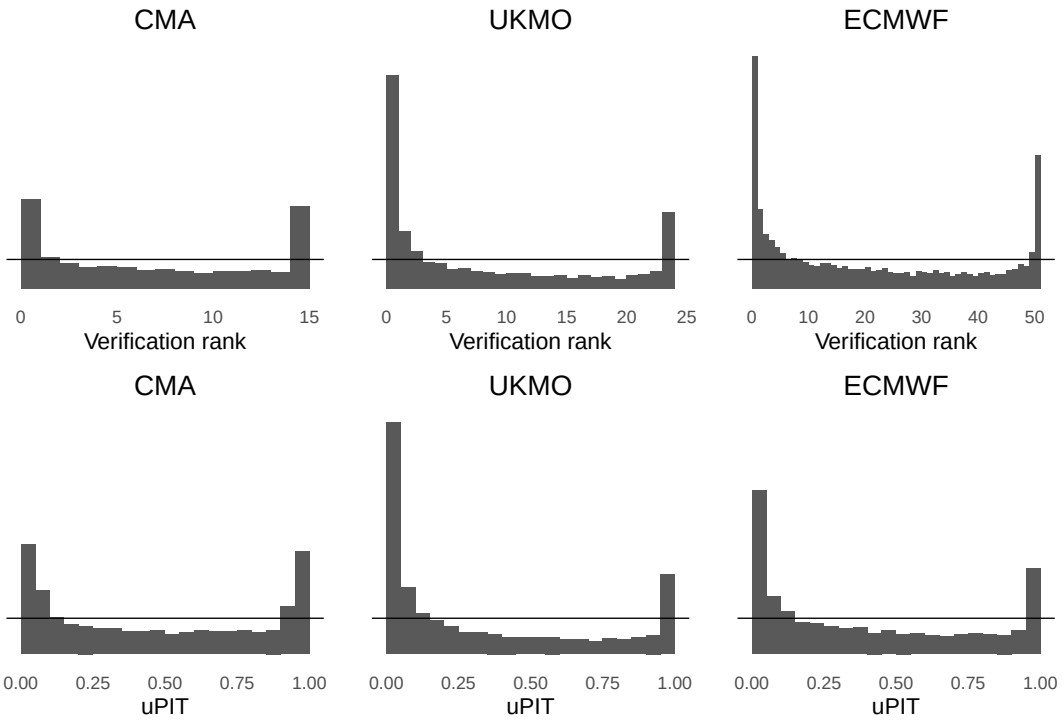


Figure 2.1: Verification rank (top row) and uPIT (bottom row) histograms for exemplary precipitation forecasts from three ensemble systems of size 14 (CMA), 23 (UKMO), and 50 (ECMWF).

for Medium-Range Weather Forecasts (ECMWF) with 50 members for 24-hour accumulated precipitation in West Sahel in 2013. Detailed information on these ensembles is provided in Table 4.1 and Section 4.1. While all three ensembles are underdispersive, the visual impression suggests that CMA is the least and ECMWF the most underdispersive ensemble.

The uPIT histograms in the bottom row of Figure 2.1 reveal that the judgement of CMA being the least underdispersed ensemble is correct. However, the assessment of the underdispersion of ECMWF was mislead by the high number of ensemble members and its underdispersion is actually less pronounced than that of UKMO.

Often, probability forecasts for binary events are of particular interest. In these settings, the CDF-valued random quantity F can be identified with a forecast $p \in [0, 1]$, representing the probability of a positive outcome $Y = 1$, in that

$$F(y) = (1 - p)\,\mathbb{1}(y \geq 0) + p\,\mathbb{1}(y \geq 1),$$

where $\mathbb{1}(A)$ shall here and in the following denote the indicator function being one if A holds and zero else. For precipitation forecasts, the probability of precipitation (PoP) as well as the probability of precipitation accumulations exceeding given amounts are of interest. A probabilistic forecast p for a binary event Y is

conditionally calibrated or *reliable* if

$$\mathbb{Q}(Y = 1 \,|\, p) = p \quad \text{almost surely.} \tag{2.5}$$

In applications, the reliability of a probabilistic forecast p is assessed in *calibration curves* or *reliability diagrams* (e.g., Murphy and Winkler, 1977; Bröcker and Smith, 2007). Given a set of forecast-observation pairs

$$\{(p_i, y_i) \,|\, i = 1, \ldots, n\}$$

representing a sample of size n from the joint distribution of p and Y, a reliability diagram plots forecast probabilities on the abscissa against conditional event frequencies on the ordinate. Specifically, a partition $0 = x_0 < x_1 < \cdots < x_N = 1$ is applied such that in each bin $[x_i, x_{i+1})$ at least a pre-specified number of forecasts is present. One then plots the empirical estimate of the conditional probability $\mathbb{Q}(Y = 1 \,|\, p \in [x_j, x_{j+1}))$ given by

$$\frac{\#\{i \,|\, p_i \in [x_j, x_{j+1}) \,, Y_i = 1\}}{\#\{i \,|\, p_i \in [x_j, x_{j+1})\}}$$

against the arithmetic center of the bin, i.e. $(x_j + x_{j+1})/2$. Bröcker and Smith (2007) note that although using the arithmetic center of a bin is very common, it has the clear disadvantage that reliable forecasts can appear unreliable. As this is not the case for the empirical mean forecast per bin, we rely on the latter instead. Deviations from the expected diagonal indicate a lack of forecast reliability and different types of forecast misspecifications can be identified by the shape of the reliability diagram (see, e.g., Wilks, 2011).

The sharpness of a forecast refers to the concentration of a predictive distribution and is a property of the forecast only. Similarly, resolution describes the ability of probability forecasts for binary outcomes to issue predictions that deviate from the unconditional probability $\mathbb{Q}(Y = 1)$.

2.3 Proper scoring rules

For the comparative assessment of forecast quality, we rely on proper scoring rules that assess calibration and sharpness simultaneously (Gneiting and Raftery, 2007; Wilks, 2011) and encourage honest and careful forecasting.

Let $\mathcal{F}$ denote a generic convex class of probability distributions F on the outcome space $\Omega_Y = \mathbb{R}$. A *scoring rule* is a mapping

$$\mathrm{S} : \mathcal{F} \times \mathbb{R} \to \mathbb{R} \cup \{\infty\} \tag{2.6}$$

that assigns a score based on the predictive distribution $F \in \mathcal{F}$ and the observation $y \in \mathbb{R}$. Typically, one assumes scoring rules to be negatively oriented and calls them *proper relative to the class* $\mathcal{F}$ if the expected score function is well-defined and

$$\mathbb{E}_{Y \sim G} \mathrm{S}(G, Y) \leq \mathbb{E}_{Y \sim G} \mathrm{S}(F, Y) \tag{2.7}$$

holds for all distributions $F, G \in \mathcal{F}$. Here, $\mathbb{E}_{Y\sim G}\mathrm{S}(G, Y)$ denotes the expected score a forecast G attains for an observation Y that is distributed as G. Strict propriety is attained if (2.7) holds with equality if and only if $F = G$.

The main benefit of propriety is that it implicitly enforces honest and careful forecasting. If a forecaster believes that the observation follows a distribution G, then G is a, and in case of strict propriety the, best forecast she can issue in order to minimize her expected score. This property is crucial, and the use of improper scoring rules can lead to misguided inferences about predictive performance as noted by Gneiting and Raftery (2007), Gneiting (2011), and Hilden and Gerds (2014).

As scoring rules summarize the predictive performance of probabilistic forecasts, they allow to assess and rank competing forecasts based on their mean scores for a given test set (Gneiting and Raftery, 2007). For probabilistic precipitation forecasts, Scheuerer (2014) argues convincingly that as forecasting precipitation is highly challenging, a small number of suboptimal forecasts should not have a too strong influence on the mean score. Additionally, the scoring rule has to accommodate forecasts consisting of a discrete component for the probability of no precipitation and a continuous component for positive accumulation amounts.

These considerations favor the *continuous ranked probability score* (CRPS; Matheson and Winkler, 1976; Gneiting and Raftery, 2007), namely

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} [F(x) - \mathbb{1}(x \geq y)]^2 \, \mathrm{d}x, \tag{2.8}$$

which is a strictly proper scoring rule relative to the class $\mathcal{F}$ of probability distributions with finite first moment. Gneiting and Raftery (2007) show that the CRPS admits the representation

$$\mathrm{CRPS}(F, y) = \mathbb{E}_F|X - y| - \frac{1}{2}\,\mathbb{E}_F|X - X'|. \tag{2.9}$$

where X and X' are independent copies of a random variable with distribution function F and finite first moment. Equation (2.9) implies that the CRPS is measured in the same unit as the forecast and is reasonably robust against outliers. While the application of the CRPS was often hindered by the lack of closed form expressions for parametric CDFs, many closed form expressions have been derived in recent years and render the CRPS also a computationally efficient choice (Gneiting and Raftery, 2007; Friederichs and Thorarinsdottir, 2012; Jordan et al., 2018).

The PoP as an essential component of any probabilistic precipitation forecast as well as exceedance probabilities above pre-defined thresholds can be evaluated by the *Brier score* (BS; Brier, 1950). For a probabilistic forecast F and an event threshold t it is given by

$$\mathrm{BS}_t(F, y) = (\mathbb{1}(y \leq t) - F(t))^2. \tag{2.10}$$

Clearly, the BS is a strictly proper scoring rule and the CRPS is the integral of the BS over all possible threshold values t. Besides the BS, many other choices are possible to evaluate probabilistic forecasts for binary events, and different choices may result in different forecast rankings.

Savage (1971) proved that subject to weak regularity conditions a scoring rule for a probability forecast p and a binary event y is proper if it can be expressed as

$$\mathrm{S}(p, y) = \phi(y) - \phi(p) - \phi'(p)(y - p), \tag{2.11}$$

where the function ϕ is convex with subgradient ϕ'. The BS arises in the case $\phi(t) = t^2$. Ehm et al. (2016) showed that in this setting every proper scoring rule admits a representation

$$\mathrm{S}(p, y) = \int_0^1 \mathrm{S}_\theta(p, y)\, \mathrm{d}H(\theta) \tag{2.12}$$

in terms of *elementary scores* or losses S_θ, namely,

$$\mathrm{S}_\theta(p, y) = \begin{cases} \theta, & y = 0, p > \theta, \\ 1 - \theta & y = 1, p \leq \theta, \\ 0, & \text{otherwise}, \end{cases} \tag{2.13}$$

and a non-negative measure H. The elementary scores can be interpreted economically, in that they reflect the cost incurred by optimal decision strategies. Given a probabilistic forecast p for a binary event y, we need to predict if it will happen or not. If correct decisions do not incur any costs, a false alarm carries cost θ, and a missed event has cost $1 - \theta$ for some $\theta \in (0, 1)$, an optimal strategy is to predict that the event will happen when $p > \theta$, and to predict that it will not happen when $p \leq \theta$.[1] Hence, θ can also be interpreted in terms of the cost-loss ratio of the decision problem (Murphy, 1977).

For evaluation purposes, Ehm et al. (2016) advocate the use of *Murphy diagrams* which display, for each forecast considered, the mean elementary score as a function of $\theta \in (0, 1)$. If a forecast receives a lower elementary score than another for every θ, it is preferable for any decision maker, and receives lower scores under just any proper scoring rule. Ehm et al. (2016) introduce for such settings the concept of forecast dominance in the Murphy diagram sense.

In many types of applications, ROC curves are popular graphical tools for the assessment of the discrimination ability of forecasts in binary prediction problems. In contrast to proper scoring rules, which assess the actual value of a forecast in decision making, ROC curves are insensitive to (any lack of) reliability and, therefore, reflect potential skill and value only (Wilks, 2011, p. 346). In Chapter 3, we study ROC curves in detail and analyze the relationship of forecast dominance in the Murphy diagram and ROC curve sense for conditionally calibrated forecasts.

[1] When $p = \theta$, either action can be taken.

2.4 Consistent scoring functions

While probabilistic forecasts are superior to point forecasts, many practical situations require single-valued point forecasts for a variety of reasons, ranging from tradition and reporting requirements to decision making (Gneiting, 2011). While one can easily transform a probabilistic forecast into a deterministic one, it is unclear how to select the "right" point forecast without any further guidance.

It is therefore necessary to specify a priori a scoring function that will be used for evaluation, and thus encourage forecasters to issue the optimal point forecast or Bayes act, or to request directly a specific functional of the forecast distribution (Gneiting, 2011). A loss or score function is called consistent for a given functional if the Bayes act corresponds to the respective functional of the forecasting distribution.

Commonly used scoring functions for the evaluation of a point forecast x for an observation y are the absolute error $\mathrm{AE}(x, y) = |x - y|$ and the squared error $\mathrm{SE}(x, y) = (x - y)^2$. It is well known that the Bayes act $\hat{x} = \arg\min \mathbb{E}_F \mathrm{S}(x, Y)$ for the AE is the median of F, while it is the mean of F for squared error. In Chapters 5 and 6, we rely on the AE to evaluate median forecasts and note that the CRPS collapses to the AE if the forecast is deterministic, as is immediate from equation (2.9).

3 Receiver Operating Characteristic (ROC) curves

In this chapter, we focus on the evaluation of the predictive ability of real-valued markers or features for binary outcomes. In particular, we introduce the concept of ROC curves and derive fundamental properties of ROC curves. We distinguish raw ROC diagnostics and ROC curves, establish the equivalence between ROC curves and CDFs on the unit interval and elucidate the crucial role of concavity in interpreting and modeling ROC curves. These results support a subtle shift of paradigms in the statistical modeling of ROC curves, which we view as curve fitting. We introduce the flexible two-parameter beta family for fitting CDFs to empirical ROC curves, derive the large sample distribution of the minimum distance estimator and currently develop software in `R` for estimation and testing, including both asymptotic and Monte Carlo based inference. In a range of empirical examples the beta family and its three- and four-parameter ramifications that allow for straight edges fit better than the classical binormal model, particularly under the vital constraint of the fitted curve being concave. Throughout Chapter 3, we closely follow Gneiting and Vogel (2018).

3.1 Introduction

Through all realms of science and society, the assessment of the predictive ability of real-valued markers or features for binary outcomes is of critical importance. To give but a few examples, biomarkers are used to diagnose the presence of cancer or other diseases, NWP systems aid in the prediction of extreme precipitation events, judges need to assess recidivism in convicts, in information retrieval documents, such as websites, are to be classified as signal or noise, banks use customers' particulars to assess credit risk, financial transactions are to be classified as fraud or no fraud, and email messages are to be identified as spam or legitimate. In these and myriads of similar settings, ROC curves are key tools in the evaluation of the predictive ability of covariates, markers or features (Egan et al., 1961; Swets, 1973, 1988; Zweig and Campbell, 1993; Fawcett, 2006). Figure 3.1 documents the astonishing rise in the use of ROC curves in the scientific literature. In 2017, nearly 8,000 papers were published that use ROC curves, up from less than 50 per year through 1990 and less than 1,000 papers annually through 2002.

A ROC curve is simply a plot of the hit rate against the false alarm rate across the range of thresholds for the real-valued marker or feature at hand. Specifically, consider the joint distribution $\mathbb{Q}$ of the pair (X, Y), where the covariate,

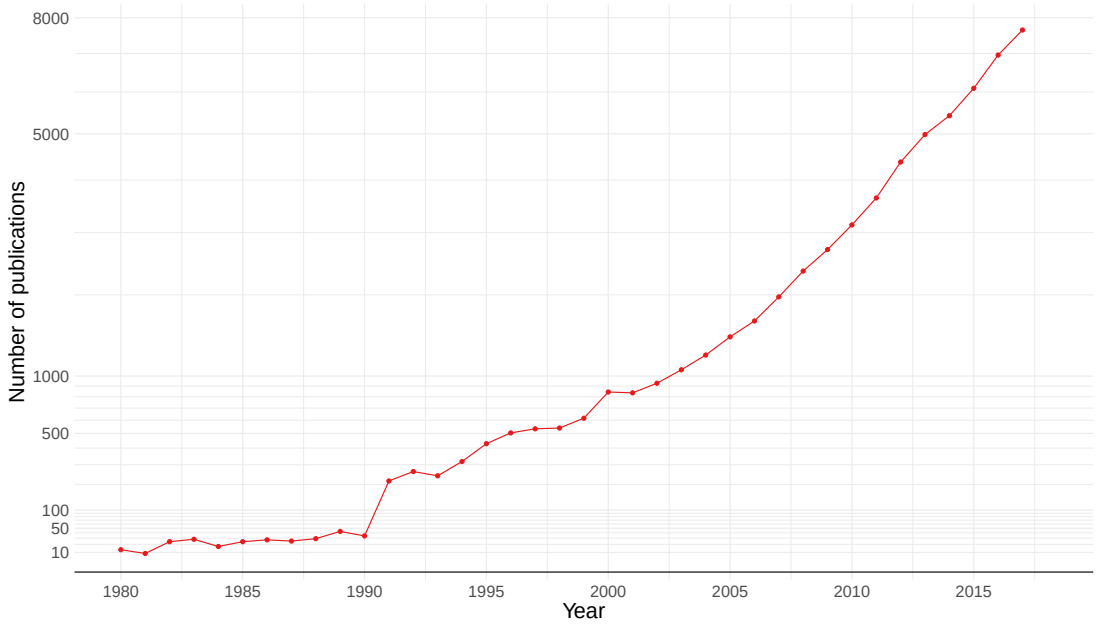


Figure 3.1: Number of publications per year resulting from a Web of Science topic search for the terms "receiver operating characteristic" or "ROC" on 24 August 2018. Note the square root scale on the vertical axis, which suggests quadratic growth.

marker or feature X is real-valued, and the event Y is binary, with the implicit understanding that higher values of X provide stronger support for the event to materialize $(Y = 1)$. The joint distribution $\mathbb{Q}$ of (X, Y) is characterized by the *prevalence* $\pi_1 = \mathbb{Q}(Y = 1) \in (0, 1)$ along with the conditional CDFs

$$F_1(x) = \mathbb{Q}(X \leq x \,|\, Y = 1) \qquad \text{and} \qquad F_0(x) = \mathbb{Q}(X \leq x \,|\, Y = 0).$$

Any threshold value x can be used to predict a positive outcome $(Y = 1)$ if $X > x$ and a negative outcome $(Y = 0)$ if $X \leq x$, to yield a classifier with *hit rate* (HR),[1]

$$\mathrm{HR}(x) = \mathbb{Q}(X > x \,|\, Y = 1) = 1 - F_1(x),$$

and *false alarm rate* (FAR),

$$\mathrm{FAR}(x) = \mathbb{Q}(X > x \,|\, Y = 0) = 1 - F_0(x).$$

The term *raw ROC diagnostic* refers to the set-theoretic union of the points of the form $(\mathrm{FAR}(x), \mathrm{HR}(x))'$ in the unit square. The *ROC curve* is a linearly

[1]Terminologies abound and differ markedly between communities. Some researchers talk of ROC as *relative operating characteristic*; see, e.g., Swets (1973) and Mason and Graham (2002). The hit rate has also been referred to as *probability of detection* (POD), *recall*, *sensitivity*, or *true positive rate* (TPR). The false alarm rate is also known as *probability of false detection* (POFD), *fall-out*, or *false positive rate* (FPR) and equals one minus the *specificity*, *selectivity*, or *true negative rate* (TNR). For an overview, see `https://en.wikipedia.org/wiki/Precision_and_recall#Definition_(classification_context)`, accessed 21 August 2018.

Table 3.1: Proposed terminology for the potential predictive strength of a feature based on the AUC value.

AUC	Descriptor
> 0.99	nearly perfect
$0.95 - 0.99$	very strong
$0.85 - 0.95$	strong
$0.75 - 0.85$	substantial
$0.65 - 0.75$	moderate
$0.50 - 0.65$	weak
≤ 0.50	abysmal

interpolated raw ROC diagnostic and therefore also a point set that may or may not admit a direct interpretation as a function. However, if F_1 and F_0 are continuous and strictly increasing, the raw ROC diagnostic and the ROC curve can be identified with a function R, where $R(0) = 0$,

$$R(p) = 1 - F_1(F_0^{-1}(1-p)) \quad \text{for} \quad p \in (0,1), \tag{3.1}$$

and $R(1) = 1$. High hit rates and low false alarm rates are desirable, so the closer the ROC curve gets to the upper left corner of the unit square the better. The area under the ROC curve (AUC) is a widely used measure of the potential predictive value of a feature (Hanley and McNeil, 1982, 1983; DeLong et al., 1988; Bradley, 1997), admitting an appealing interpretation as the probability of a marker value drawn from F_1 being higher than a value drawn independently from F_0. Table 3.1 proposes terminology for the description of the strength of the potential value in terms of AUC.

In data analytic practice, the measure $\mathbb{Q}$ is the empirical distribution of a sample $(x_i, y_i)_{i=1}^n$ of real-valued features x_i and corresponding binary observations y_i. To generate a ROC curve in this setting, it suffices to consider the unique values of $x_1, \ldots, x_n$ and the respective false alarm and hit rates. The resulting raw ROC diagnostic is interpolated linearly to yield an empirical ROC curve, as illustrated in Figure 3.2 on examples from the biomedical (Etzioni et al., 1999; Sing et al., 2005; Robin et al., 2011) and meteorological (Vogel et al., 2018) literatures. Based on AUC and the terminology in Table 3.1, the predictor strength is moderate in the example from Robin et al. (2011), substantial for the data from Etzioni et al. (1999) and Vogel et al. (2018), and strong in the example from Sing et al. (2005). Arguably, the immense popularity of empirical ROC curves and AUC across the scientific literature stems from their ease of implementation and interpretation in concert with a wide range of desirable properties, such as invariance under strictly increasing transformations of a feature.

The remainder of this chapter is organized as follows. Section 3.2 establishes some fundamental theoretical results. We formalize the distinction between raw ROC diagnostics and ROC curves, demonstrate an equivalence between ROC curves and CDFs, and elucidate the special role of concavity in the interpretation

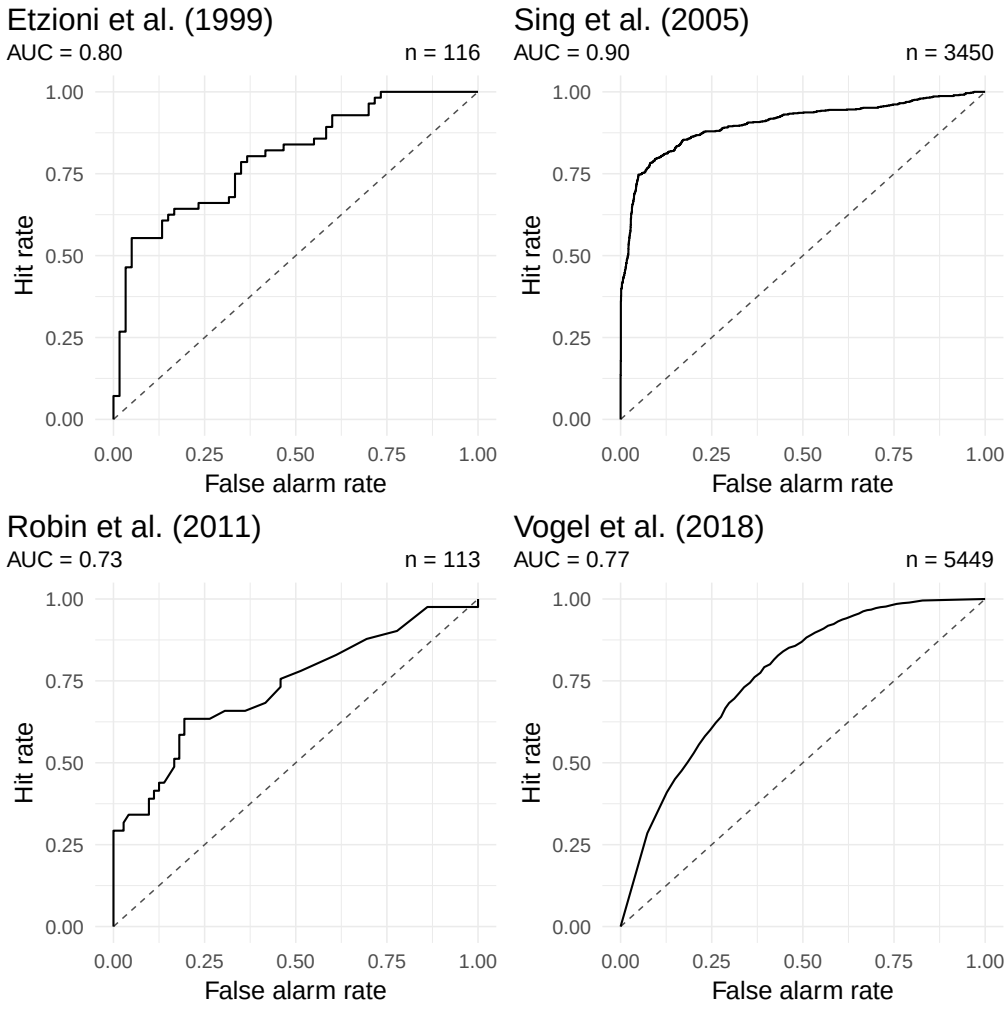


Figure 3.2: Examples of empirical ROC curves.

and modeling of ROC curves. In Section 3.3 we introduce the flexible yet parsimonious two-parameter beta model, which uses the CDFs of beta distributions to model ROC curves, and we discuss estimation and testing based on empirical ROC curves, including both asymptotic and Monte Carlo based approaches. We derive the asymptotic distribution of the minimum distance estimator in general parametric settings, and specialize to both the beta family and the classical binormal model. Section 3.4 returns to our empirical examples, of which we present detailed analyses, with the beta family and its natural three- and four-parameter extensions that allow for straight edges in the ROC curve fitting better than the binormal model, particularly under the concavity constraint. Section 3.5 presents the `betaROC` package and this chapter closes with a discussion in Section 3.6. Proofs of a more technical character are deferred to Appendices 3.A-3.D.

3.2 Fundamental properties of ROC curves

Consider the bivariate random vector (X, Y) where X is a real-valued predictor, *covariate*, *feature*, or *marker*, and Y is the binary response. We refer to the joint distribution of (X, Y) as $\mathbb{Q}$. Let $\pi_1 = \mathbb{Q}(Y = 1) \in (0, 1)$ and $\pi_0 = 1 - \pi_1 = \mathbb{Q}(Y = 0)$, and let $F_1(x) = \mathbb{Q}(X \leq x \,|\, Y = 1)$, $F_0(x) = \mathbb{Q}(X \leq x \,|\, Y = 0)$, and

$$F(x) = \mathbb{Q}(X \leq x) = \pi_0 F_0(x) + \pi_1 F_1(x)$$

denote the conditional and marginal cumulative distribution functions (CDFs) of X, respectively. Furthermore, we let $F_0(x-) = \lim_{x' \uparrow x} F_0(x')$.

We use column vectors to denote points in the Euclidean plane, and given any $(a, b)' \in \mathbb{R}^2$ we write $(a, b)'_{(1)} = a$ and $(a, b)'_{(2)} = b$ for the respective coordinate projections.

3.2.1 Raw ROC diagnostics and ROC curves

In this common setting ROC diagnostics concern the points of the form $(\mathrm{FAR}(x), \mathrm{HR}(x))'$, where $\mathrm{FAR}(x) = 1 - F_0(x)$ is the *false alarm rate* and $\mathrm{HR}(x) = 1 - F_1(x)$ the *hit rate* at the threshold value $x \in \mathbb{R}$. Formally, the *raw ROC diagnostic* for the random vector (X, Y) and the bivariate distribution $\mathbb{Q}$ is the point set

$$R^* = \left\{ \begin{pmatrix} 1 - F_0(x) \\ 1 - F_1(x) \end{pmatrix} : x \in \mathbb{R} \right\} \tag{3.2}$$

within the unit square. Clearly, the bivariate distribution $\mathbb{Q}$ of (X, Y) is characterized by F_0, F_1, and any of the two marginal distributions. In contrast, the raw ROC diagnostic along with a single marginal does not characterize $\mathbb{Q}$, due to the well known invariance of ROC diagnostics under strictly increasing transformations of X and shifts in the prevalence of the binary outcome (Fawcett, 2006). However, the raw ROC diagnostic along with both marginal distributions determines $\mathbb{Q}$.

Theorem 3.1. *The joint distribution $\mathbb{Q}$ of (X, Y) is characterized by the raw ROC diagnostic and the marginal distributions of X and Y.*

Proof. The mapping $g : [0, 1]^2 \to [0, 1]$ defined by

$$(a, b)' \mapsto (1 - a)\pi_0 + (1 - b)\pi_1$$

induces a bijection between the raw ROC diagnostic R^* and the range of F. Therefore, it suffices to note that $\mathbb{Q}(X \leq x, Y \leq y) = 0$ for $y < 0$,

$$\begin{aligned} \mathbb{Q}(X \leq x, Y \leq y) &= F_0(x)\,\pi_0 \\ &= F(x) - (1 - \mathrm{HR}(x))\,\pi_1 \\ &= F(x) - (1 - g^{-1}_{(2)}(F(x)))\,\pi_1 \end{aligned}$$

for $y \in [0, 1)$, and $\mathbb{Q}(X \leq x, Y \leq y) = F(x)$ for $y \geq 1$. □

Table 3.2: Ordered marker values $x_1 < x_2 < \cdots < x_7$, binary observations, and FAR and HR at the respective threshold for the example in Figure 3.3.

X	$< x_1$	x_1	x_2	x_3	x_4	x_5	x_6	x_7	$> x_7$
Y		0	1	0, 0	0, 0, 1	0, 1, 1	1	1	
FAR $\times$ 6	6	5	5	3	1	0	0	0	0
HR $\times$ 6	6	6	5	5	4	2	1	0	0

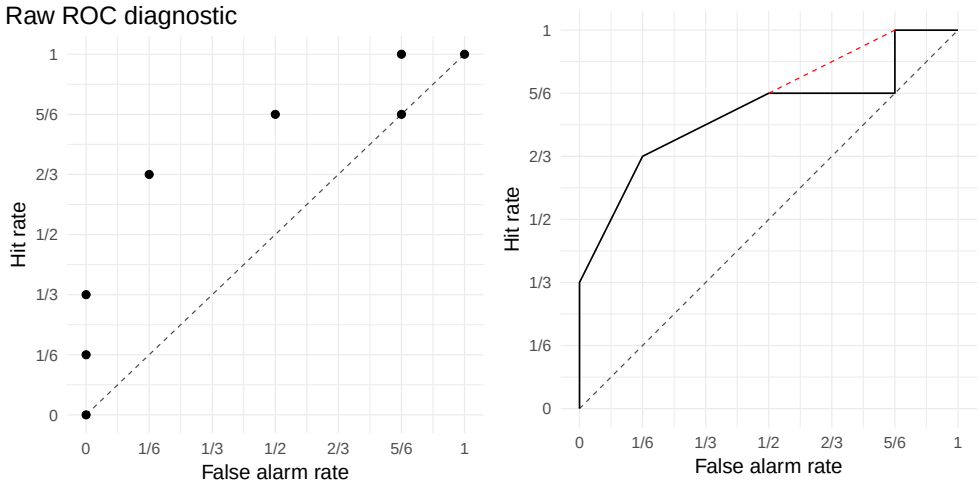


Figure 3.3: Raw ROC diagnostic (left) and corresponding empirical ROC curve (right) for the marker in Table 3.2. The broken red line completes the concave hull of the empirical ROC curve.

Briefly, a ROC curve is obtained from the raw ROC diagnostic by linear interpolation. Formally, the *full ROC diagnostic* or *ROC curve* is the point set

$$R = \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right\} \cup R^* \cup \{L_x : x \in \mathbb{R}\} \cup \left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\} \tag{3.3}$$

within the unit square, where

$$L_x = \left\{ \alpha \begin{pmatrix} 1 - F_0(x-) \\ 1 - F_1(x-) \end{pmatrix} + (1-\alpha) \begin{pmatrix} 1 - F_0(x) \\ 1 - F_1(x) \end{pmatrix} : \alpha \in [0,1] \right\}$$

is a possibly degenerate, nondecreasing line segment. The choice of linear interpolation to complete the raw ROC diagnostic into the ROC curve (3.3) is natural and persuasive, as the line segment L_x represents randomized combinations of the classifiers associated with its end points. In particular, linear interpolation allows for a fair and direct comparison between continuous, discrete, and ordinal features. Empirical ROC curves based on samples, as illustrated in Figure 3.2, fit this framework, as they arise in the special case where $\mathbb{Q}$ is an empirical measure. We illustrate the transition from the raw ROC diagnostic to the ROC curve in Figure 3.3 using the toy data set from Table 3.2, where there are twelve observations and seven unique marker values.

The raw ROC diagnostic can be recovered from the ROC curve and the two marginal distributions, as the mapping g in the proof of Theorem 3.1 induces a bijection between the raw ROC diagnostic and the range of F that can be expressed in terms of π_1 and π_0. From this simple fact the following result is immediate.

Corollary 3.2. *The joint distribution $\mathbb{Q}$ of (X, Y) is characterized by the ROC curve and the marginal distributions of X and Y.*

In this sense, ROC curves and raw ROC diagnostics assume roles similar to those of copulas (e.g., Nelsen, 2006) with the difference that ROC curves are defined in terms of conditional distributions, whereas copulas operate on marginal distributions.

Given a ROC curve R, an obvious task is to find CDFs F_0 and F_1 that realize R. For a particularly simple and appealing construction, let F_0 be the CDF of the uniform distribution on the unit interval, and take F_1 to be F_{NI}, defined as $F_{\mathrm{NI}}(x) = 0$ for $x \leq 0$,

$$F_{\mathrm{NI}}(x) = 1 - R_+(1 - x) \quad \text{for} \quad x \in (0, 1), \tag{3.4}$$

and $F_{\mathrm{NI}}(x) = 1$ for $x \geq 1$, where the function $R_+ : (0, 1) \to [0, 1]$ is induced by the ROC curve at hand, in that

$$R_+(x) = \inf \{b : (a, b)' \in R,\ a \geq x\}.$$

In anticipation of its repeated use in subsequent sections, we refer to this specific realization of a ROC curve R, in which F_0 is standard uniform and F_1 is taken to be F_{NI} in (3.4), as the *natural identification*.

Remarkably, the natural identification applies even when the feature X is discrete or ordinal. Nevertheless, the statistical models and methods that we introduce in Section 3.3 target the case of a continuous marker or feature.

3.2.2 Concave ROC curves

We proceed to elucidate the critical role of concavity in the interpretation and modeling of ROC curves.[2] Its significance is well known and has been alluded to in monographs, such as by Egan (1975, p. 35), Pepe (2003, p. 71), and Zhou et al. (2011, p. 40). Nevertheless, we are unaware of any rigorous treatment in the extant literature. To address this omission, we distinguish and analyse regular and discrete settings. Unified treatments are feasible but considerably technical, and we leave them to future work.

[2] Again, terminologies differ between communities. In machine learning, concave ROC curves are typically referred to as *convex* (e.g., Fawcett, 2006), whereas the psychological and biomedical literatures call them *proper* (Egan, 1975, Section 2.6; Zhou et al., 2011, Section 2.7.3). The usage in this thesis is in accordance with well established, commonly used terminology in the mathematical sciences.

In the *regular setting* we suppose that F_1 and F_0 have continuous, strictly positive Lebesgue densities f_1 and f_0 in the interior of an interval, which is their common support. For every x in the interior of the support, we can define the *likelihood ratio*,

$$\mathrm{LR}(x) = \frac{f_1(x)}{f_0(x)},$$

and the *conditional event probability*,

$$\mathrm{CEP}(x) = \mathbb{Q}(Y = 1 \,|\, X = x) = \frac{\pi_1 f_1(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)}.$$

We demonstrate the equivalence of the following three conditions:

(a) The ROC curve is concave.

(b) The likelihood ratio is nondecreasing.

(c) The conditional event probability is nondecreasing.

Theorem 3.3. *In the regular setting statements (a), (b), and (c) are equivalent.*

Proof. In the regular setting the ROC curve can be identified with a function $R : [0,1] \to [0,1]$, where $R(p)$ is defined as in (3.1) for $p \in (0,1)$. If the ROC curve is concave then clearly the function R is concave as well, and so its derivative $R'(p)$ is nonincreasing in $p \in (0,1)$. However, the slope $R'(p)$ equals the likelihood ratio $\mathrm{LR}(x)$ at a certain value x that decreases with p, which establishes the equivalence of (a) and (b). Furthermore,

$$\mathrm{LR}(x) = \frac{\pi_0}{\pi_1} \frac{\mathrm{CEP}(x)}{1 - \mathrm{CEP}(x)},$$

and the function $c \mapsto c/(1-c)$ is nondecreasing in $c \in (0,1)$, which yields the equivalence of (b) and (c). □

Next we consider the *discrete setting* in which the support of the feature X is a finite or countably infinite set. This setting includes, but is not limited to, the case of empirical ROC curves, as illustrated in Figure 3.2. For every x in the discrete support of X, we can define the *likelihood ratio*,

$$\mathrm{LR}(x) = \begin{cases} \mathbb{Q}(X = x \,|\, Y = 1) \,/\, \mathbb{Q}(X = x \,|\, Y = 0) & \text{if} \quad \mathbb{Q}(X = x \,|\, Y = 0) > 0, \\ \infty, & \text{if} \quad \mathbb{Q}(X = x \,|\, Y = 0) = 0, \end{cases}$$

and the *conditional event probability*,

$$\mathrm{CEP}(x) = \mathbb{Q}(Y = 1 \,|\, X = x).$$

In Appendix 3.A we prove the following direct analogue of Theorem 3.3.

Theorem 3.4. *In the discrete setting statements (a), (b), and (c) are equivalent.*

The critical role of concavity in the interpretation and modeling of ROC curves stems from the monotonicity condition (c) on the conditional event probability, which is at the very heart of the approach and needs to be invoked to justify the construction of just any raw ROC diagnostic or ROC curve. In the medical literature Hilden (1991) notes that "some authors do seem to overlook the concavity problem" and Pesce et al. (2010) argue that "direct use of a decision variable" with a non-concave ROC curve "must be considered irrational" and "unethical when applied to medical decisions". Similar considerations apply in the vast majority of applications of ROC curves.

Fortunately, there are straightforward ways of restricting attention to concave ROC curves and the associated classifiers. Generally, randomization can be used to generate classifiers with concave ROC curves from features with non-concave ones (Fawcett, 2006; Pesce et al., 2010). The regular setting serves to supply theoretical models that can be fit to empirical ROC curves, such as the classical binormal model or our new beta model, and the parameters in these models can be restricted suitably to guarantee concavity, as we discuss in Section 3.3. Empirical ROC curves typically fail to be concave, as illustrated in Figure 3.2. However, they can readily be morphed into their concave hull, by subjecting the marker or feature at hand to the pool-adjacent violators (PAV: Ayer et al., 1955; De Leeuw et al., 2009) algorithm, thereby converting it into an isotonic, calibrated probabilistic classifier (Lloyd, 2002; Fawcett and Niculescu-Mizil, 2007). For example, for the toy data in Table 3.3 the PAV algorithm assigns the conditional event probability $p_1 = 0$ to x_1, the value $p_2 = \frac{1}{3}$ to x_2, x_3, and x_4, the value $p_3 = \frac{2}{3}$ to x_5, and the value $p_4 = 1$ to x_6 and x_7. The ROC curve for this isotonic and calibrated probabilistic classifier is the concave hull of the ROC curve for the original marker, as shown in Figure 3.3.

3.2.3 Equivalence of ROC curve and Murphy diagram dominance for conditionally calibrated forecasts

Based on the invariance of ROC curves under strictly increasing transformations of the feature, ROC curves evaluate potential, rather than real, skill of a predictor. In contrast, Murphy diagrams evaluate real skill of predictions as noted in Chapter 2. If predictions are conditionally calibrated, then the notions of potential and real skill coincide and so should the interpretation of Murphy diagrams and ROC curves. In particular, if a forecast has a ROC curve that is everywhere to the top left of the ROC curve of another forecast and therefore dominates it in the ROC curve sense, then it should also receive lower elementary scores, introduced in (2.13), in the Murphy diagram across all thresholds and as such dominate it in the Murphy diagram sense. Even before the introduction of Murphy diagrams by Ehm et al. in 2016, Wilks (2011, p. 346) noted with reference to Krzysztofowicz and Long (1990) that

> "On the other hand, when forecasts underlying ROC diagrams are correctly calibrated, dominance of one ROC curve over another (i.e., one curve lying entirely above and to the left of another) implies statistical sufficiency for the dominating forecasts, so that these will be of greater use for all rational forecast users."

With the Murphy diagram indicating greater economic value by lower elementary scores, Wilks' statement suggests the equivalence of ROC curve and Murphy diagram dominance in case of (conditionally) calibrated forecast. In the following, the equivalence of the notions of forecast dominance in the ROC curve and in the Murphy diagram sense is proven for calibrated forecasts. To this end, the next two theorems introduce characterizations of ROC curves and Murphy diagrams in the calibrated forecast setting.

Theorem 3.5. *Let X be a discrete or absolutely continuous, conditionally calibrated probability forecast for Y. Then the ROC curve is completely determined by the marginal distribution of X.*

Proof. We start by proving the claim for discrete forecasts X, so that the support of X is a finite or countably infinite, ordered set of two or more points $x_i \in [0,1]$, indexed by consecutive integers $i \in I$ such that $x_i < x_j$ if $i < j$. In the case of a finite set, we assume that it is at least of cardinality two.[3] Denote by $f(x_i) = \mathbb{Q}(X = x_i) > 0$ for $i \in I$ the probability of X attaining the value x_i. Then

$$\pi_0 = \mathbb{Q}(Y = 0) = \sum_{i \in I} \mathbb{Q}(Y = 0 \,|\, X = x_i)\, \mathbb{Q}(X = x_i) = \sum_{i \in I} (1 - x_i) f(x_i),$$

as the conditional calibration of X guarantees $\mathbb{Q}(Y = 0 \,|\, X = x_i) = 1 - x_i$. Similarly, $\pi_1 = \mathbb{Q}(Y = 1)$ is also determined by f. The Bayes theorem implies

$$\begin{aligned} f_0(x_i) &= \mathbb{Q}(X = x_i \,|\, Y = 0) = \mathbb{Q}(Y = 0 \,|\, X = x_i)\, \mathbb{Q}(X = x_i)\, \frac{1}{\mathbb{Q}(Y = 0)} \\ &= (1 - x_i)\, f(x_i)\, \pi_0^{-1}, \\ f_1(x_i) &= \mathbb{Q}(X = x_i \,|\, Y = 1) = \mathbb{Q}(Y = 1 \,|\, X = x_i)\, \mathbb{Q}(X = x_i)\, \frac{1}{\mathbb{Q}(Y = 1)} \\ &= x_i\, f(x_i)\, \pi_1^{-1} \end{aligned}$$

and the conditional distributions $F_0(x_i) = \mathbb{Q}(X \leq x_i \,|\, Y = 0)$ and $F_1(x_i) = \mathbb{Q}(X \leq x_i \,|\, Y = 1)$ are completely determined by f. As F_0 and F_1 characterize the associated ROC curve, the statement follows. In case of a continuous forecast distribution f on $[0,1]$, the result follows with suitable technical adaptions in an analogous way. □

[3] The raw ROC diagnostic in case of a set of cardinality one is represented by one single combination of hit and false alarm rate situated on the chance diagonal for conditionally calibrated forecasts.

Theorem 3.6. *Let X be a discrete or absolutely continuous, conditionally calibrated probability forecast for Y. Then its Murphy diagram is completely determined by the marginal distribution of X.*

Proof. The Murphy diagram plots the expected elementary score $\mathbb{E}\mathrm{S}_\theta(X, Y)$ introduced in (2.13) as a function of the threshold $\theta \in [0, 1]$ and can be expressed as

$$\begin{aligned}\mathbb{E}\mathrm{S}_\theta(X, Y) &= \theta\, \mathbb{Q}(Y = 0, X > \theta) + (1 - \theta)\, \mathbb{Q}(Y = 1, X \leq \theta) \\ &= \theta\, \mathbb{Q}(X > \theta \,|\, Y = 0)\, \mathbb{Q}(Y = 0) + (1 - \theta)\, \mathbb{Q}(X > \theta \,|\, Y = 1)\, \mathbb{Q}(Y = 1) \\ &= \theta\, (1 - F_0(\theta))\, \pi_0 + (1 - \theta)\, F_1(\theta)\, \pi_1.\end{aligned}$$

As the distribution of X determines F_0, F_1, π_0, and π_1 in case of a conditionally calibrated forecast, the statement follows. □

Definition 3.7. *Let X_1 and X_2 be probability forecasts for a binary event Y. Then X_1 dominates X_2 in the Murphy diagram sense if*

$$\mathbb{E}\mathrm{S}_\theta(X_1, Y) \leq \mathbb{E}\mathrm{S}_\theta(X_2, Y) \tag{3.5}$$

holds for all $\theta \in [0, 1]$. Forecast X_1 strictly dominates X_2 if (3.5) holds and $\mathbb{E}S_\theta(X_1, Y) < \mathbb{E}S_\theta(X_2, Y)$ for some $\theta \in (0, 1)$.

Definition 3.8. *Let X_1 and X_2 be probability forecasts for a binary event Y. Forecast X_1 dominates forecast X_2 in the ROC curve sense if*

$$F_{\mathrm{NI},1}(x) \leq F_{\mathrm{NI},2}(x) \tag{3.6}$$

holds for all $x \in [0, 1]$. Here, $F_{NI,1}$ and $F_{NI,2}$ are the CDFs obtained by the natural identification of the ROC curves R_1 and R_2 corresponding to forecasts X_1 and X_2. Furthermore, X_1 strictly dominates X_2 if (3.6) holds and $F_{NI,1}(x) < F_{NI,2}(x)$ for some $x \in (0, 1)$.

Theorem 3.9. *For discrete or absolutely continuous, conditionally calibrated probability forecasts X_1 and X_2 for the binary outcome Y, (strict) dominance in the Murphy diagram sense is equivalent to (strict) dominance in the ROC curve sense.*

The proof proceeds by introducing an alternative characterization of ROC curve dominance, that allows to invoke the concavity of both ROC curves. Subsequently, reformulations yield the equivalence between (strict) ROC curve and (strict) Murphy diagram dominance. See Appendix 3.B for details of the proof.

3.2.4 An equivalence between ROC curves and probability measures

We move on to provide concise and practically relevant characterizations of ROC curves, both with and without the critical condition of concavity.

Theorem 3.10. *There is a one-to-one correspondence between ROC curves and probability measures on the unit interval. In particular, the natural identification induces a bijection between the class of the ROC curves and the class of the CDFs of probability measures on the unit interval.*

Proof. Given a ROC curve, we can remove any vertical line segments, except for the respective upper endpoints, to yield the CDF of a probability measure on the unit interval. Conversely, given the CDF of a probability measure on the unit interval, we can interpolate vertically at any jump points to obtain a ROC curve. This mapping is a bijection, and save for the symmetries in (3.4) is realized by the natural identification. □

We say that a curve C in the Euclidean plane is nondecreasing if $a_0 \leq a_1$ is equivalent to $b_0 \leq b_1$ for points $(a_0, b_0)', (a_1, b_1)' \in C$. The following result is immediate.

Corollary 3.11. *The ROC curves are the nondecreasing curves in the unit square that connect the points* $(0, 0)'$ *and* $(1, 1)'$.

We now state characterizations under the constraint of strict concavity. Analogous results hold under the slightly weaker assumption of concavity.

Theorem 3.12. *There is a one-to-one correspondence between strictly concave ROC curves and probability measures with strictly increasing Lebesgue densities on the unit interval, which is induced by the natural identification.*

Corollary 3.13. *The strictly concave ROC curves are in one-to-one correspondence to the strictly concave functions* R *on the unit interval with* $R(0) = 0$ *and* $R(1) = 1$.

Turning to methodological and applied considerations, these results support a shift of paradigms in the statistical modeling of ROC curves. In extant practice, the emphasis is on modeling the conditional distributions F_0 and F_1, such as in the ubiquitous binormal model. Our results suggest a subtle but important change of perspective, in that ROC modeling can be approached as an exercise in curve fitting,[4] with any nondecreasing curve that connects $(0, 0)'$ to $(1, 1)'$ being a permissible candidate, and parametric families of CDFs on the unit interval offering particularly attractive models, including but not limited to the beta family that we introduce in the next section.

3.3 Parametric models, estimation, and testing

The binormal model is by far the most frequently used parametric model and "plays a central role in ROC analysis" (Pepe, 2003, p. 81). Specifically, the

[4] While curve fitting approaches have been advocated before, such as by Swets (1986, p. 104, his approach (b)), they lacked theoretical support.

binormal model assumes that F_1 and F_0 are Gaussian with means $\mu_1 \geq \mu_0$ and strictly positive variances σ_0^2 and σ_1^2, respectively. We are in the regular setting of Subsection 3.2.2, and the resulting ROC curve is represented by the function $R : [0,1] \to [0,1]$ with $R(0) = 0$,

$$R(p) = \Phi(\mu + \sigma \Phi^{-1}(p)) \quad \text{for} \quad p \in (0,1), \tag{3.7}$$

and $R(1) = 1$, where Φ is the CDF of the standard normal distribution, $\mu = (\mu_1 - \mu_0)/\sigma_1 \geq 0$ is a scaled difference in expectations, and $\sigma = \sigma_0/\sigma_1$ is the ratio of the respective standard deviations. The respective area under the curve is

$$\mathrm{AUC}(\mu, \sigma) = \Phi\left(\frac{\mu}{\sqrt{1+\sigma^2}}\right).$$

For an illustration of binormal ROC curves see the left-hand panel of Figure 3.4. It is well known that a binormal ROC curve is concave only if $\sigma = 1$ or equivalently if F_0 and F_1 differ in location only. Theorem 3.14 generalizes this impossibility result to obtain concavity under different variances to location–scale-families in general.

Proposition 3.14. *Let $\mathcal{F}$ denote the class of strictly increasing CDFs on $\mathbb{R}$. For any $F \in \mathcal{F}$, set $F_0(x) = F(x)$ and $F_1(x) = F\left(\frac{x-\mu}{\sigma}\right)$ for some $\mu > 0$ and $\sigma > 0$. Then the ROC curve associated with the conditional CDFs F_0 and F_1 is non-concave whenever $\sigma \neq 1$.*

Proof. A ROC curve is non-concave if it crosses the diagonal given by $\mathrm{HR}(x) = \mathrm{FAR}(x)$ for at least one $x \in \mathbb{R}$. It holds that

$$\mathrm{HR}(x) < \mathrm{FAR}(x) \quad \Leftrightarrow \quad F_0(x) < F_1(x) \quad \Leftrightarrow \quad F_0(x) < F_0\left(\frac{x-\mu}{\sigma}\right).$$

Suppose $\sigma > 1$. Then $F_0(x) < F_1(x)$ for all $x < -\mu/(\sigma - 1)$. If $\sigma < 1$, then $F_0(x) < F_1(x)$ for all $x > \mu/(1-\sigma)$. □

Under the binormal model, concave ROC curves are necessarily symmetric with respect to the anti-diagonal in the unit square, which strongly inhibits their flexibility as illustrated in the left-hand panel of Figure 3.5.

3.3.1 The beta model

Motivated and supported by the characterization theorems of Section 3.2, we propose a curve fitting approach to the statistical modeling of ROC curves, with the two-parameter family of the cumulative distribution functions (CDFs) of beta distributions being a particularly attractive model. Specifically, consider the beta family with ROC curves represented by the function

$$R(p) = B_{\alpha,\beta}(p) = \int_0^p b_{\alpha,\beta}(q)\,\mathrm{d}q \quad \text{for} \quad p \in [0,1], \tag{3.8}$$

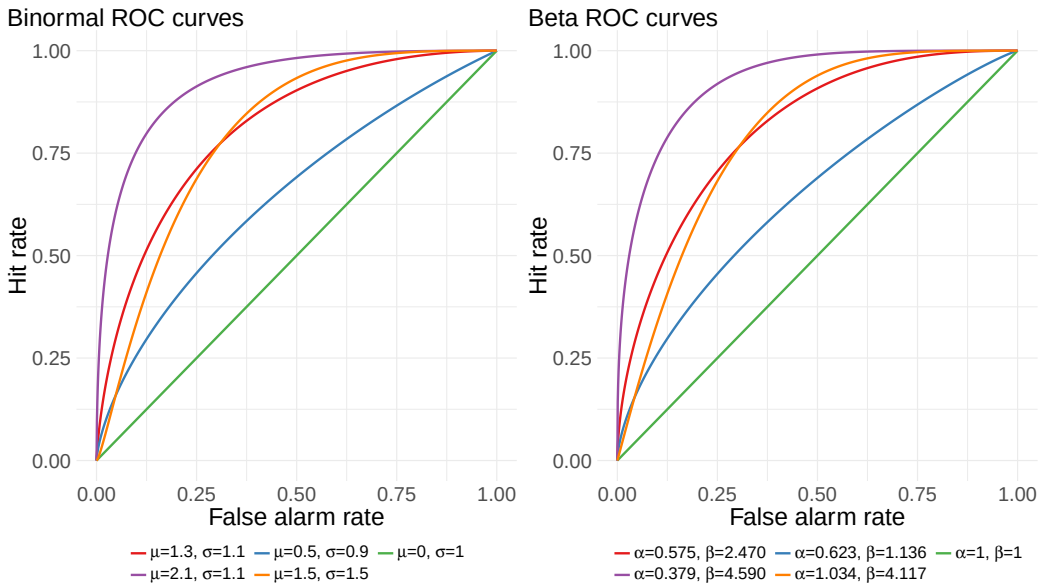


Figure 3.4: Members of the (left) binormal family and (right) beta family of ROC curves. The parameter values for the beta curves have been chosen to match the overall shape of the same-color binormal ROC curve.

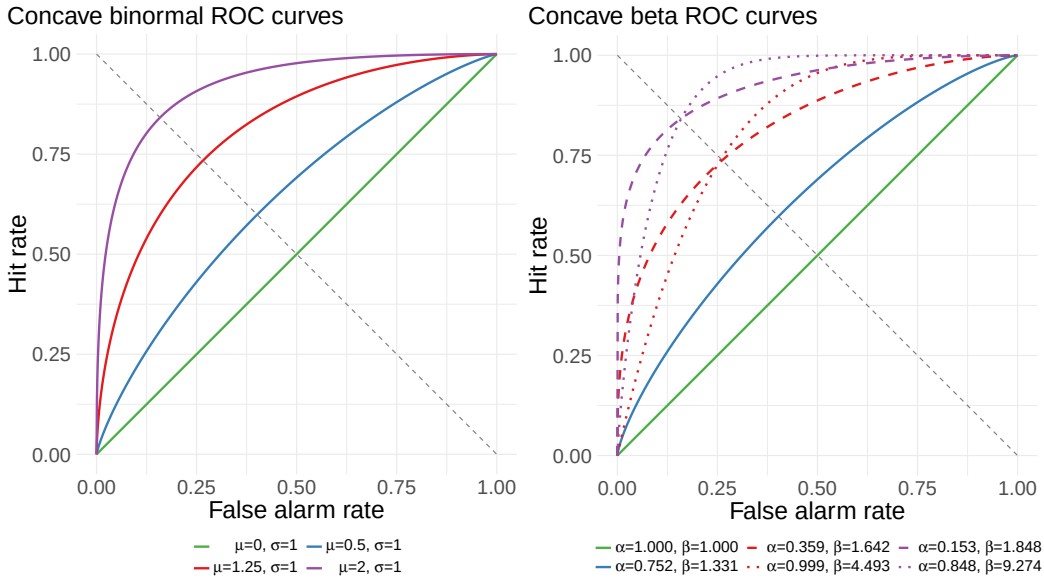


Figure 3.5: Concave members of the (left) binormal family and (right) beta family of ROC curves. The parameter values for the beta curves have been chosen to match the value of the same-color binormal ROC curve at the anti-diagonal.

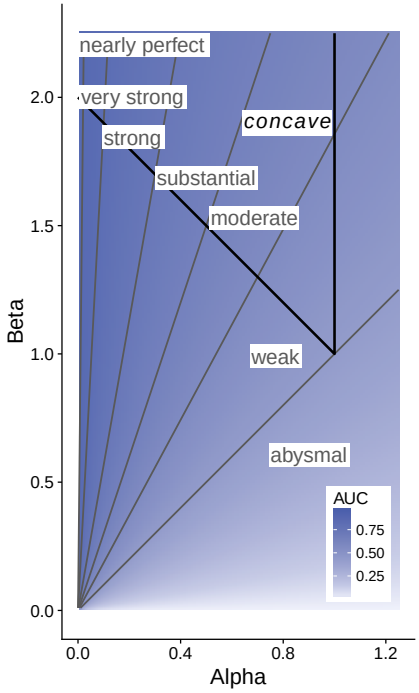


Figure 3.6: AUC value for the beta family of ROC curves. The isolines correspond to the terminology for predictor strength introduced in Table 3.1. Parameter combinations above and left of the black line yield concave ROC curves.

where $b_{\alpha,\beta}(q) \propto q^{\alpha-1}(1-q)^{\beta-1}$ is the density of the beta distribution with parameter values $\alpha > 0$ and $\beta > 0$. As illustrated in Figure 3.6 and shown in Appendix 3.C, a beta ROC curve is concave if $\alpha \leq 1$ and $\beta \geq 2-\alpha$, and its AUC value is

$$\mathrm{AUC}(\alpha, \beta) = \frac{\beta}{\alpha + \beta}.$$

In the limit as $\beta \to \infty$ we obtain the perfect ROC curve with straight edges from $(0,0)'$ to $(0,1)'$ and $(1,1)'$, corresponding to a complete separation of the supports of F_1 and F_0. While the requirement of a concave ROC curve is restrictive, the condition is much less stringent than for the binormal family, where it constrains the admissible parameter space to a single dimension. The adaptability of the beta family is illustrated in Figure 3.4, where we see that members of the beta family can match the shape of binormal ROC curves, and in Figure 3.5, where the gain in flexibility under the critical constraint of concavity is evident.

The beta family nests the time-honored one-parameter power model (Egan et al., 1961; Swets, 1986) that arises in the special case when $\beta = 1$. While the classical derivation of the power model does not readily generalize, our theoretical results justify the use of the two-parameter beta family. If even further flexibility

is desired, mixtures of beta CDFs, i.e., functions of the form

$$R_n(p) = \sum_{k=1}^{n} w_k B_{\alpha_k,\beta_k}(p) \quad \text{for} \quad p \in [0,1],$$

where $w_1, \ldots, w_n \geq 0$ with $w_1 + \cdots + w_n = 1$, $\alpha_1, \ldots, \alpha_k > 0$, and $\beta_1, \ldots, \beta_k > 0$, approximate any regular ROC curve to any desired accuracy, as demonstrated by the following result. Recall from Subsection 3.2.2 that in the regular setting the ROC curve can be identified with the function R in (3.1), where F_1 and F_0 have continuous, strictly positive Lebesgue densities f_1 and f_0 in the interior of an interval, which is their common support. A ROC curve is *regular* if it arises in this way and *strongly regular* if furthermore the derivative R' is bounded.

Theorem 3.15. *For every strongly regular ROC curve R there is a sequence of mixtures of beta CDFs that converges uniformly to R.*

The proof of this result relies on Bernstein's probabilistic approach to the Weierstrass theorem (Levasseur, 1984) and is deferred to Appendix 3.C.

3.3.2 Minimum distance estimation

For the parametric estimation of ROC curves for continuous markers various methods have been proposed, including maximum likelihood (Dorfman and Alf, 1969; Metz et al., 1998; Zou and Hall, 2000), approaches based on generalized linear models (Pepe, 2000), and minimum distance estimation (Hsieh and Turnbull, 1996), as reviewed at book length by Pepe (2003), Krzanowski and Hand (2009), and Zhou et al. (2011).

Maximum likelihood techniques face a conceptual challenge, in that ROC curves do not determine the joint distribution of the marker and the binary event. Here we pursue the minimum distance estimator, which is much in line with our curve fitting approach.

We assume a parametric model in the regular setting of Subsection 3.2.2, where now the ROC curve depends on a parameter $\theta \in \Theta \subseteq \mathbb{R}^k$. Specifically, we suppose that for each $\theta \in \Theta$ the ROC curve is represented by a smooth function

$$R(p;\theta) = 1 - F_{1,\theta}(F_{0,\theta}^{-1}(1-p)) \quad \text{for} \quad p \in (0,1),$$

where $F_{1,\theta}$ and $F_{0,\theta}$ admit continuous, strictly positive densities $f_{1,\theta}$ and $f_{0,\theta}$ in the interior of an interval, which is their common support. We also require that the true parameter value θ_0 is in the interior of the parameter space Θ, where the derivative

$$R'(p;\theta) = \frac{\partial R(p;\theta)}{\partial p} = \frac{f_{1,\theta}(F_{0,\theta}^{-1}(1-p))}{f_{0,\theta}(F_{0,\theta}^{-1}(1-p))}$$

exists and is finite for $p \in (0,1)$, and where the partial derivative $R_{(i)}(p;\theta)$ of $R(p;\theta)$ with respect to component i of the parameter vector $\theta = (\theta_1, \ldots, \theta_k)'$ exists and is continuous for $i = 1, \ldots, k$ and $p \in (0,1)$.

We adopt the asymptotic scenario of Hsieh and Turnbull (1996) where at sample size n there are n_0 and $n_1 = n - n_0$ independent draws from $F_{0,\theta}$ and $F_{1,\theta}$ with corresponding binary outcomes of zero and one, respectively, and where $\lambda_n = n_0/n_1$ converges to some $\lambda \in (0, \infty)$ as $n \to \infty$. For $\theta \in \Theta$ we define the difference process

$$\xi_n(p; \theta) = \hat{R}_n(p) - R(p; \theta),$$

where the function $\hat{R}_n(p)$ represents the empirical ROC curve. The minimum distance estimator $\hat{\theta}_n = (\hat{\theta}_1, \ldots, \hat{\theta}_k)'_n$ then satisfies

$$\|\xi_n(\cdot; \hat{\theta}_n)\| = \min_{\theta \in \Theta} \|\xi_n(\cdot; \theta)\|,$$

where $\|\xi_n(\cdot; \theta)\| = (\int_0^1 \xi_n(p; \theta)^2 \, \mathrm{d}p)^{1/2}$ is the standard L_2-norm. If n is large, $\hat{\theta}_n$ exists and is unique with probability approaching one (Millar, 1984) and so we follow the extant literature in ignoring issues of existence and uniqueness.

The minimum distance estimator has a multivariate normal limit distribution in this setting, as suggested by the asymptotic result of Hsieh and Turnbull (1996) that under the usual $\sqrt{n}$ scaling the difference process $\xi_n(p; \theta)$ has limit

$$W(p; \theta) = \sqrt{\lambda} \, B_1(R(p; \theta)) + R'(p; \theta) \, B_2(p) \tag{3.9}$$

at $\theta = \theta_0$, where B_1 and B_2 are independent copies of a Brownian bridge. In Appendix 3.D we review the specifics of the convergence to the limit process (3.9) and combine results of Millar (1984) and Hsieh and Turnbull (1996) to show the following result.

Theorem 3.16. *In the above setting the minimum distance estimator $\hat{\theta}_n$ satisfies*

$$\sqrt{n} \, (\hat{\theta}_n - \theta_0) \to \mathcal{N}(0, C^{-1} A C^{-1}) \tag{3.10}$$

as $n \to \infty$, where the matrices A and C have entries

$$A_{ij} = \int_0^1 \int_0^1 R_{(i)}(s; \theta_0) \, K(s, t; \theta_0) \, R_{(j)}(t; \theta_0) \, \mathrm{d}s \, \mathrm{d}t, \quad C_{ij} = \int_0^1 R_{(i)}(s; \theta_0) \, R_{(j)}(s; \theta_0) \, \mathrm{d}s \tag{3.11}$$

for $i, j = 1, \ldots, k$, respectively, and where

$$\begin{aligned} K(s, t; \theta_0) = \lambda(\min\{R(s; \theta_0), R(t; \theta_0)\} - R(s; \theta_0) R(t; \theta_0)) \\ + R'(s; \theta_0) R'(t; \theta_0) (\min\{s, t\} - st). \end{aligned} \tag{3.12}$$

is the covariance function of the process $W(p; \theta)$ in (3.9) at $\theta = \theta_0$.

Corollary 3.17. *In the above setting,*

$$\sqrt{n} \, (\mathrm{AUC}(\hat{\theta}_n) - \mathrm{AUC}(\theta_0)) \to \mathcal{N}(0, G C^{-1} A C^{-1} G'), \tag{3.13}$$

where G is the gradient of the mapping $\theta \mapsto AUC(\theta)$ at $\theta = \theta_0$.

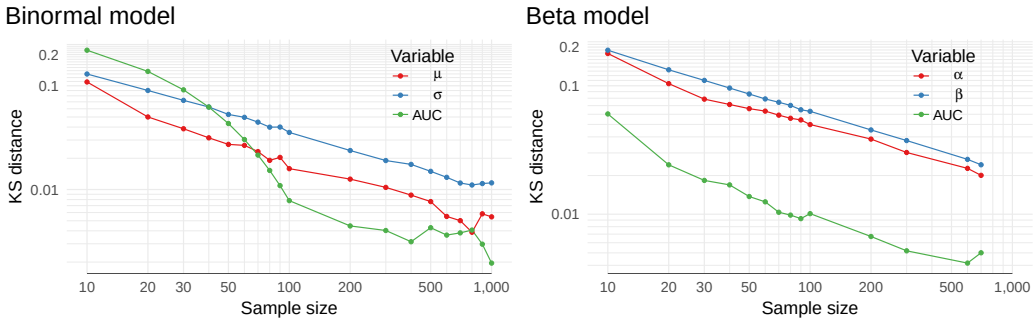


Figure 3.7: Convergence to the Gaussian limit in (3.10) and (3.13). For each sample size n, we show the Kolmogorov–Smirnov (KS) distance between the (scaled) empirical distribution of the minimum distance estimate of the quantity of interest, as described in Subsection 3.3.2, and the respective Gaussian limit. Left: Binormal model at $(\mu_0, \sigma_0) = (0.50, 1.00)$. Right: Beta model at $(\alpha_0, \beta_0) = (0.67, 2.00)$.

Both the binormal and the beta model satisfy the assumptions for these results, which allows for asymptotic inference about the model parameters and the AUC, by plugging in $\hat{\theta}_n$ for θ_0 in the expressions for the asymptotic covariances. For the binormal model (3.7) we have $\theta = (\mu, \sigma)$, $R_{(\mu)}(p;\theta) = \varphi(\mu+\sigma\Phi^{-1}(p))$, $R_{(\sigma)}(p;\theta) = \Phi^{-1}(p)\varphi(\mu + \sigma\Phi^{-1}(p))$, and $R'(p;\theta) = \sigma\varphi(\mu + \sigma\Phi^{-1}(p))/\varphi(\Phi^{-1}(p))$, where φ is the standard normal density, so that the integrals in (3.11) can readily be evaluated numerically. The gradient in (3.13) equals $G = (\varphi(\mu_0/\sqrt{1+\sigma_0^2})/(1+\sigma_0^2)^{1/2}, -\mu_0\sigma_0\varphi(\mu_0/\sqrt{1+\sigma_0^2})/(1+\sigma_0^2)^{3/2})$. Hsieh and Turnbull (1996) consider binormal ordinal dominance curves, which interchange the roles of F_1 and F_0 relative to ROC curves, and after reparameterization we recover their results. However, the formula for the covariance function $K(s,t;\theta)$ in the first displayed equation on page 39 in Hsieh and Turnbull (1996) is incompatible with our equation (3.12) and incorrect, as it is independent of s and t and therefore constant. Under the beta model (3.8) we have $\theta = (\alpha, \beta)$ and $R'(p;\theta) = b_{\alpha,\beta}(p)$. While closed form expressions for the partial derivatives of $R(p;\theta)$ with respect to α and β exist, they are difficult to evaluate, and we approximate them with finite differences. The gradient in (3.13) equals $G = (-\beta_0/(\alpha_0+\beta_0)^2, \alpha_0/(\alpha_0+\beta_0)^2)$.

Figure 3.7 illustrates the convergence to the Gaussian limit distributions in Theorem 3.16 and Corollary 3.17 in a Monte Carlo study. For each sample size n considered we let $\lambda_n = 1$, draw $N = 200,000$ samples of size n, find the associated empirical ROC curves, and compute the respective minimum distance estimates $\hat{\theta}_{n,1}, \ldots, \hat{\theta}_{n,N}$. Then we consider the empirical distribution of the components of $\{\sqrt{n}\,(\hat{\theta}_{n,i} - \theta_0) : i = 1, \ldots, N\}$ and $\{\sqrt{n}\,(\mathrm{AUC}(\hat{\theta}_{n,i}) - \mathrm{AUC}(\theta_0)) : i = 1, \ldots N\}$. Even at moderate sample sizes n, the scaled empirical distributions are close to their Gaussian limits.

3.3.3 Testing goodness-of-fit and other hypotheses

We move on to discuss testing. A natural hypothesis to be addressed is whether a given parametric model fits the data at hand. In contrast to existing methods that are based on AUC and focus on the binormal model (Zou et al., 2005), we propose a simple Monte Carlo test that applies to any parametric model $\mathcal{C}$. For example, $\mathcal{C}$ could be the full binormal, the concave binormal, the full beta, or the concave beta family. While we describe the procedure for minimum distance estimates and the L_2-distance, it applies equally to other estimates and other distance measures.

Given a dataset of size n with n_0 instances where the binary outcome is zero and $n_1 = n - n_0$ instances where it is one, our goodness-of-fit test proceeds as follows. We use the notation of Subsection 3.3.2 and denote the number of Monte Carlo replicates by M.

1. Fit a model from class $\mathcal{C}$ to the empirical ROC curve for the data at hand, to yield the minimum distance estimate θ_{data}. Compute d_{data} as the L_2-distance between the fitted and the empirical ROC curve.

2. For $m = 1, \ldots, M$,
   a) draw a sample of size n under θ_{data}, with n_0 and n_1 instances from $F_{0,\theta_{\text{data}}}$ and $F_{1,\theta_{\text{data}}}$ and associated binary outcomes of zero and one, respectively,
   b) fit a model from class $\mathcal{C}$ to the empirical ROC curve, to yield the minimum distance estimate, and
   c) compute d_m as the L_2-distance between the fitted and the empirical ROC curve.

3. Find a p-value based on the rank of d_{data} when pooled with $d_1, \ldots, d_M$. Specifically, $p = (\#\{i = 1, \ldots, M : d_{\text{data}} \leq d_i\} + 1)/(M + 1)$.

Under the null hypothesis of the ROC curve being generated by a random sample within class $\mathcal{C}$ the Monte Carlo p-value is very nearly uniformly distributed, as is readily seen in simulation experiments (not reported on here).

Parametric tests of the equality of ROC curves and AUC values can be based on the limit distributions in Theorem 3.16 and Corollary 3.17 in the usual way. Under an identifiable model the hypothesis of two ROC curves being equal is the same as the hypothesis of the respective parameters being the same. Therefore, the limit in (3.10) allows for a customary chi square test of the equality of ROC curves from independent samples, based on the squared norm of the normalized difference between the two estimates of the parameter vector, as proposed by Metz and Kronman (1980) in the case of maximum likelihood estimates under the binormal model. Similarly, the limit in (3.13) justifies a z-test for the equality of the AUC values, based on the normalized difference between the two parametric estimates of the AUC. We illustrate the use of these tests in the subsequent section and provide software for their implementation in the case of independent

samples. For paired, dependent samples, correlations between the estimates need to be accounted for, a task to be addressed in future work. As an alternative, nonparametric tests have been developed in the extant literature (Hanley and McNeil, 1983; DeLong et al., 1988; Venkatraman and Begg, 1996; Venkatraman, 2000; Mason and Graham, 2002).

3.4 Empirical examples

We return to the empirical ROC curves in Figure 3.2 and present basic information about the underlying datasets in Table 3.3. In the dataset from Etzioni et al. (1999), the negative logarithm of the ratio of free to total prostate-specific antigen (PSA) two years prior to diagnosis in serum from patients later found to have prostate cancer is compared to age-matched controls. The datasets from Sing et al. (2005, Figure 1a) and Robin et al. (2011, Figure 1) are prominent examples in the widely used `ROCR` and `pROC` packages in `R`. They concern a score from a linear support vector machine (SVM) trained to predict the usage of human immunodeficiency virus (HIV) coreceptors, and the S100β biomarker as it relates to a binary clinical outcome, respectively. The dataset from Vogel et al. (2018, Figure 6d) considers PoP forecasts from the ECMWF NWP ensemble system for the binary event of precipitation occurrence within the next 24 hours at meteorological stations in the West Sahel region in northern tropical Africa. This dataset is discussed in Chapter 5 and Figures 5.1 and 5.5.

Figure 3.8 shows binormal and beta ROC curves fitted to the empirical ROC curves, both in the unrestricted case and under the constraint of concavity. The respective unrestricted and restricted minimum distance estimates, the fit in terms of the L_2-distance to the empirical ROC curve, and the p-value from the goodness-of-fit test in Subsection 3.3.3 with $M = 999$ Monte Carlo replicates, are given in Table 3.3. In the unrestricted case, the binormal and beta fits are visually nearly indistinguishable. The fitted binormal ROC curves fail to be concave and change markedly when concavity is enforced. For the beta ROC curves, the differences between restricted and unrestricted fits are less pronounced, and in the example from Vogel et al. (2018) the unrestricted fit is concave. Generally, in the constrained case the improvement in the fit under the more flexible beta model as compared to the classical binormal model is substantial.

The theoretical results in Subsection 3.3.2 allow for asymptotic inference about the model parameters. We illustrate this in Figure 3.9 for the unrestricted beta fit for the dataset from Etzioni et al. (1999). In addition to showing confidence ellipsoids, we indicate and separate concave and non-concave fits. If we seek to complement the minimum distance estimate with pointwise confidence bands for the ROC curve, we can sample from the inferred distribution for the model parameters and display the envelope of the respective ROC curves, as exemplified in Figure 3.11.

A closer look at the empirical ROC curves for the biomedical data from Etzioni et al. (1999), Sing et al. (2005), and Robin et al. (2011) in Figures 3.2 and

Table 3.3: Basic information about the datasets and minimum distance estimates under the unrestricted and concave binormal and beta models for the ROC curves in Figures 3.2 and 3.8. Fit is in terms of the L_2-distance to the empirical ROC curve, and the p-value is from the goodness-of-fit test of Subsection 3.3.3.

Dataset	Etzioni et al. (1999)	Sing et al. (2005)	Robin et al. (2011)	Vogel et al. (2018)
Binary outcome	prostate cancer	coreceptor usage	clinical outcome	precipitation
Feature	antigen ratio	SVM predictor	S100β concentr.	NWP forecast
Sample size	116	3450	113	5449
Binormal model				
unrestricted (μ, σ)	(1.05, 0.78)	(1.58, 0.65)	(0.75, 0.72)	(1.13, 1.22)
fit	0.043	0.019	0.033	0.008
p-value	0.106	0.001	0.561	0.032
concave (μ, σ)	(1.22, 1.00)	(2.05, 1.00)	(0.91, 1.00)	(0.99, 1.00)
fit	0.056	0.039	0.060	0.031
p-value	0.138	0.001	0.147	0.001
Beta model				
unrestricted (α, β)	(0.34, 1.32)	(0.15, 1.44)	(0.36, 0.96)	(0.79, 2.57)
fit	0.042	0.023	0.032	0.006
p-value	0.117	0.001	0.620	0.187
concave (α, β)	(0.38, 1.62)	(0.17, 1.83)	(0.51, 1.49)	(0.79, 2.57)
fit	0.045	0.025	0.050	0.006
p-value	0.196	0.001	0.204	0.171

3.8 reveals a striking commonality, in that the curves show vertical and/or horizontal straight edges. From the definition of the raw ROC characteristic (3.2) it is evident that straight edges correspond to marker values that may allow for deterministic class attribution, as illustrated in the back-to-back histograms in Figure 3.10. Importantly, straight edges might convey critical information from a subject matter perspective, such as in medical diagnoses, where straight edges in ROC curves correspond to particularly high or low marker values that might identify individuals as healthy or diseased beyond doubt.

Under the beta family the statistical modeling of straight edges is straightforward. Specifically, we can generalize the two-parameter model (3.8) to a four-parameter beta family, where

$$R(p) = \gamma + (1 - \gamma)\, B_{\alpha,\beta}\Big(\frac{p}{\delta}\Big) \quad \text{for} \quad p \in (0, 1], \tag{3.14}$$

which allows for a vertical straight edge that connects the coordinate origin $(0, 0)'$ to the point $(0, \gamma)'$, and a horizontal straight edge that connects the points $(\delta, 1)'$ and $(1, 1)'$ within the ROC curve. Three-parameter subfamilies with a single type of straight edge arise if we fix $\delta = 1$ and let $\gamma \in [0, 1]$ vary, or fix $\gamma = 0$ and consider $\delta \in (0, 1]$, respectively. While the subfamily with $\delta = 1$ being fixed has a direct analogue under the binormal model, there is no natural way of adapting

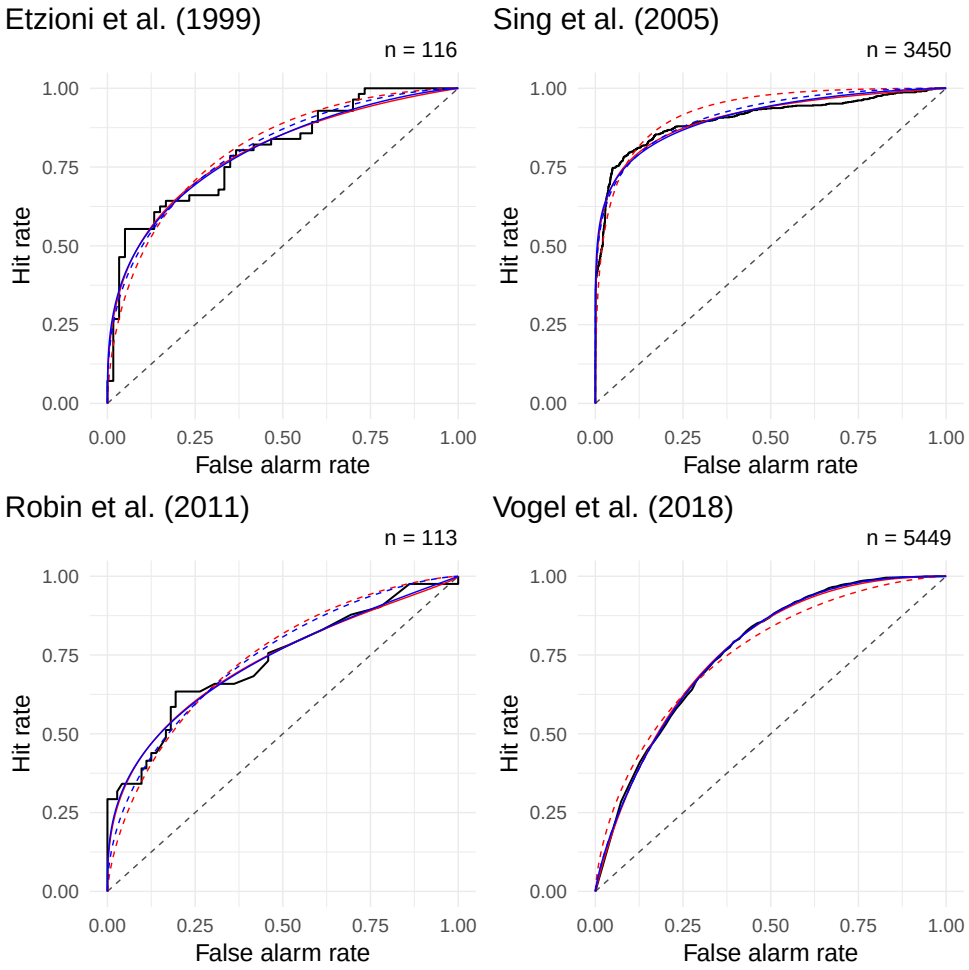


Figure 3.8: Fitted binormal (red) and beta (blue) ROC curves in the unrestricted (solid) and concave (dashed) case for the datasets from Figure 3.2 and Table 3.3.

the subfamily with $\gamma = 0$ being fixed or the four-parameter family in (3.14) to the binormal case.

To be clear, we do *not* advocate uncritical routine use of the four-parameter family in (3.14) and the respective three-parameter subfamilies. However, we *do* recommend that in any specific application researchers check for straight edges in empirical ROC curves, and assess on the basis of substantive expertise whether or not they ought to be modeled. Visual tools such as the back-to-back histograms for the conditional distributions in Figure 3.10 can assist in this assessment. For illustration, the back-to-back histograms might suggest that we fit the three-parameter model with $\gamma = 0$ being fixed to the data from Etzioni et al. (1999) and the three-parameter model with $\delta = 1$ being fixed to the data from Sing et al. (2005) and Robin et al. (2011). While in the first two cases the three-parameter fits are nearly identical to the fits under the two-parameter beta model, the three-parameter extension yields a substantially improved fit for the data from Robin

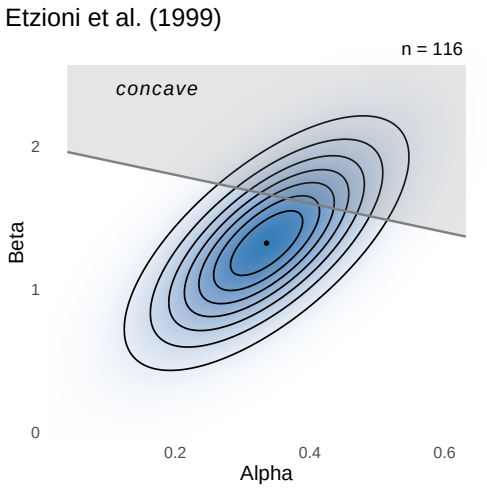


Figure 3.9: Asymptotic inference under the unrestricted beta model for the data from Etzioni et al. (1999). The confidence ellipses are at level $1/8, 2/8, \ldots, 7/8$, respectively.

et al. (2011) as illustrated in the lower right panel of Figure 3.10. The constrained minimum distance estimate for (α, β, γ) is $(0.70, 1.30, 0.24)$ with L_2-distance 0.029 to the empirical ROC curve. For comparison, under the two-parameter concave beta model the estimate for (α, β) is $(0.51, 1.49)$ with L_2-distance 0.050.

Finally, we take another look at the meteorological data from Vogel et al. (2018). Here it is obvious from the scientific context in weather prediction that the above three- and four-parameter extensions are irrelevant. While the data introduced and analyzed in Table 3.3 and Figures 3.2 and 3.8 concern PoP forecasts over the West Sahel region, Vogel et al. (2018) consider the East Sahel region as well.[5] The respective empirical ROC curves are shown in Figure 3.11 along with the constrained two-parameter beta fit and parametric 95% pointwise confidence bands. The p-value for the goodness-of-fit test of Subsection 3.3.3 is 0.168 for West Sahel and 0.057 for East Sahel. Our parametric tests for equality of AUC values and ROC curves yield p-values of 0.633 and 0.015, whereas the nonparametric tests of DeLong et al. (1988) and Venkatraman (2000) result in p-values of 0.616 and 0.089, respectively.

3.5 `R` package `betaROC`

While studying ROC curves, we have developed software for the statistical programming language `R` (R Core Team, 2018) that is currently prepared for release as `betaROC` package. A preliminary version of the package is available online at `https://github.com/PeterVogel1991/betaROC`. The aim of the `betaROC` package is to provide user-friendly tools to study and analyze the predictive ability of

[5] See Chapter 5 for the respective study, and Figure 5.1 for the location and extent of both Sahelian regions.

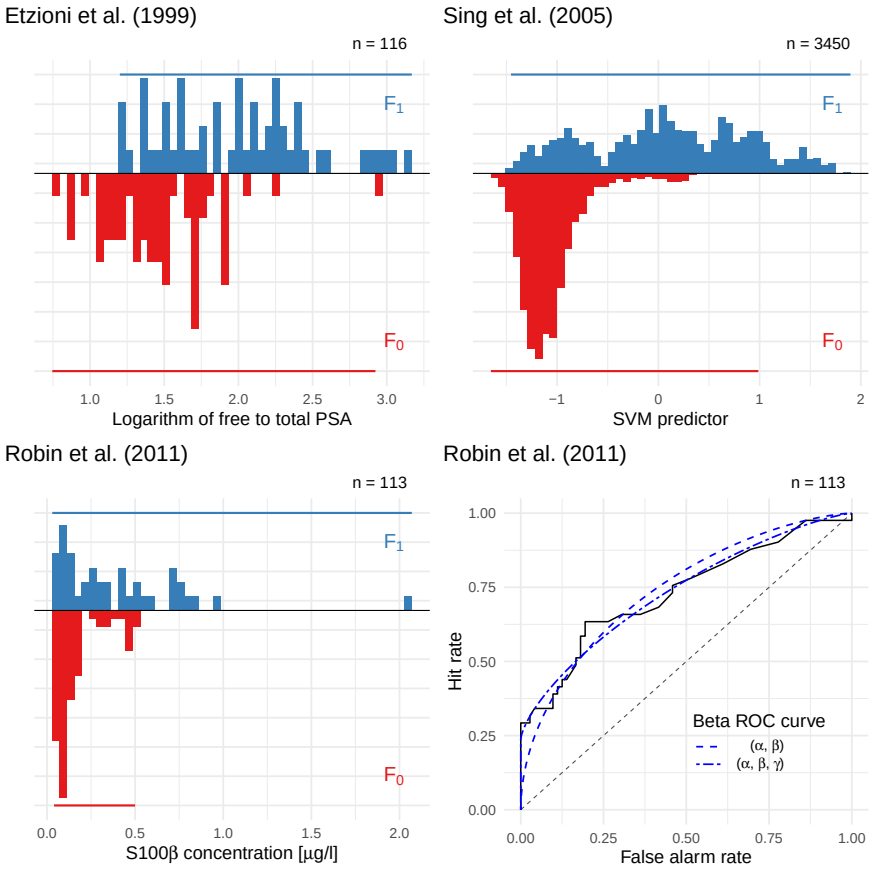


Figure 3.10: Histograms for the conditional distributions for data in Table 3.3, and concave two- and three-parameter beta ROC curves fit to the data from Robin et al. (2011). The horizontal lines in the histograms extend to the convex hull of the respective support.

features, markers, and predictions for binary outcomes. Starting from the empirical data consisting of feature, marker, or prediction values and the corresponding binary observations, the `betaROC` package allows to visualize the empirical conditional distributions and their support (Figure 3.10) and to compute and visualize the raw ROC diagnostic and corresponding ROC curve (Figures 3.2 and 3.3).

For beta ROC curves, minimum distance estimates (MDEs) can be computed for the 2-parameter model (α, β), the 3-parameter models (α, β, γ) and (α, β, δ), and the full 4-parameter model $(\alpha, \beta, \gamma, \delta)$. Based on the conceptual restriction of the binormal ROC model, only the 2-parameter model (μ, σ) and the 3-parameter model (μ, σ, γ) are available. All MDE fits can be restricted to allow for concave ROC curves only.

For the estimated parameters of the beta or binormal ROC curves, the asymptotic distribution can be visualized (Figure 3.9), and the MDE fitted curves can be plotted along with the empirical ROC curves (Figures 3.4 and 3.8). Tests for goodness-of-fit are available for both two-parameter fits, but are computationally demanding. Additionally tests for the equality of two ROC curves for

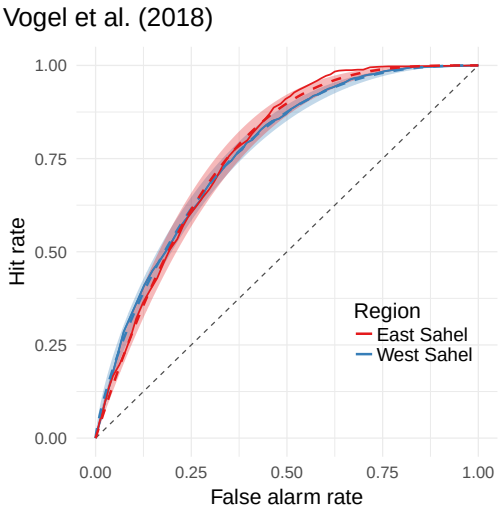


Figure 3.11: Empirical ROC curve (solid), concave beta fit (dashed), and associated pointwise 95% confidence band (shaded) for data from Vogel et al. (2018) on PoP forecasts over West (blue) and East (red) Sahel in northern tropical Africa.

unpaired data are available as are tests to check the hypothesis of equal predictive performance as measured by the AUC for unpaired data.

Other `R` packages to compute and visualize ROC curves as well as derived properties exist. Most noticeable are the `R` packages `pROC` (Robin et al., 2011) and `ROCR` (Sing et al., 2005). `ROCR` computes many performance measures for ROC curves, provides e.g. accuracy, calibration, and conditional density plots, but does not feature theoretical models or tests for the equality of ROC curves or predictive performance measures. Similarly, the `pROC` package allows to compute ROC properties such as AUC values or confidence intervals, to smooth ROC curves to obtain a binormal ROC curve and to apply non-parametric tests to empirical ROC curves. As the asymptotic distribution for binormal as well as for beta ROC curves has to the best of our knowledge not been correctly derived and computed beforehand, all tests and visualizations that rely on the asymptotic distribution have not been incorporated in any other software package. Additionally, the `betaROC` package incorporates code and data to reproduce all figures in this section except for Figures 3.7 and 3.12.

3.6 Discussion

ROC curves have been used extensively to evaluate the potential predictive value of covariates, features, or markers in binary problems in a multitude of scientific disciplines. Their appeal stems from attractive and desirable properties in this context, which include the straightforward interpretation of ROC curves in terms of attainable operating conditions (i.e., hit and false alarm rates), their invariance under strictly increasing transformations of the feature and shifts in prevalence,

and the interpretation of AUC as the probability of a marker value drawn from F_1 being higher than a value drawn independently from F_0. We emphasize that ROC curves and AUC values "should be regarded as a measure of potential rather than actual skill" (Kharin and Zwiers, 2003, p. 4148) tailored to serve the purposes of variable selection and feature screening across all types of ordinal, discrete, and continuous predictor variables.[6]

Despite their ubiquitous use, our understanding of fundamental properties of ROC curves has been incomplete. The theoretical results in Section 3.2 establish an equivalence between ROC curves and the CDFs of probability measures on the unit interval, which motivates and justifies our curve fitting approach to the statistical modeling of ROC curves. Concave fits are preferred, if not essential, as they characterize the predictor variables with nondecreasing likelihood ratios and nondecreasing conditional event probabilities. The beta family (3.8) provides a particularly attractive parametric model. As compared to the classical binormal model the beta family is considerably more flexible under the constraint of concavity, and it embeds naturally into the four-parameter model (3.14) that allows for straight edges in the ROC curve. If further flexibility is sought, mixtures of beta CDFs can be fitted. With a view toward nonparametric alternatives, one might model (minus) the second derivative of a regular ROC curve, which is nonnegative under the concavity constraint.

For estimation we focus on the minimum distance approach. In the regular setting, where features are continuous, minimum distance estimates and associated parametric estimates of the AUC value are asymptotically normal. Goodness of fit and other hypotheses can be tested for based on these methods and results. In view of the critical role of concavity for the interpretation of ROC curves, an interesting and relevant question is whether or not one should subject features to the PAV algorithm (De Leeuw et al., 2009) prior to fitting a concave model. The PAV algorithm morphs the empirical ROC curve into the respective concave hull, and its use for data pre-processing in other types of shape-constrained estimation problems has been examined by Mammen (1991). The derivation of the large sample distributions in Subsection 3.3.2 is based on empirical process theory (Shorack and Wellner, 2009), and it depends on the Gaussian limit in (3.9), which does not apply under ordinal or discrete features nor when ROC curves have straight edges. We leave the derivation of large sample distributions for minimum distance estimates in these cases as well as adaptations to covariate- and time-dependent settings (Etzioni et al., 1999; Heagerty et al., 2000) to future work. Datasets and code in `R` (R Core Team, 2018) for replicating our results and implementing the proposed estimators and tests will be released soon.

[6] ROC curves and AUC values have limitations when they are used to assess the actual skill of probability forecasts, as they ignore the critical requirement of calibration (Wilks, 2011, p. 346). For evaluating the *actual* skill and value of probabilistic classifiers, proper scoring rules (Gneiting and Raftery, 2007) are a preferred tool, notably in the form of Murphy diagrams (Ehm et al., 2016). For a direct comparison of ROC curves and Murphy diagrams and a respective discussion in the context of probability forecasts see Figure 5.5 and Section 5.3.

Appendix 3.A Concave ROC curves: The discrete setting

In proving Theorem 3.4 we may assume that the support of X is a finite or countably infinite, ordered set of two or more points x_i, indexed by consecutive integers such that $x_i < x_j$ if $i < j$. In the case of a finite set we assume that it is of cardinality at least 2 and adapt the arguments in obvious ways to account for boundary effects.

Lemma 3.18. *Any of the statements in Theorem 3.4 implies that either*

(i) $\mathbb{Q}(X = x_i \,|\, Y = 0) > 0$ *for all* i*, or*

(ii) there exists an index value i^* *such that* $\mathbb{Q}(X = x_i \,|\, Y = 0) = 0$ *for all* $i \geq i^*$ *and* $\mathbb{Q}(X = x_i \,|\, Y = 0) > 0$ *for all* $i < i^*$*.*

Proof. If any of the statements in Theorem 3.4 hold and condition (i) is violated, there exists an index i such that $\mathbb{Q}(X = x_i \,|\, Y = 0) = 0\}$. Then $\mathrm{CEP}(x_i) = 1$, $\mathrm{LR}(x_i) = \infty$, and the ROC curve has a vertical straight edge away from the origin, which contradicts statements (c), (b), and (a), respectively, unless condition (ii) is satisfied. □

Proof of Theorem 3.4. In view of Lemma 3.18, it suffices to show the equivalence of the statements in Theorem 3.4 for indices i with $\mathbb{Q}(X = x_i \,|\, Y = 0) > 0$. The fact that

$$\mathrm{LR}(x_i) = \frac{\pi_0}{\pi_1} \frac{\mathrm{CEP}(x_i)}{1 - \mathrm{CEP}(x_i)}$$

along with the monotonicity of the function $c \mapsto c/(1-c)$ establishes the equivalence of (b) and (c). Furthermore, the relationship

$$\begin{aligned} \mathrm{LR}(x_i) &= \frac{\mathbb{Q}(X = x_i \,|\, Y = 1)}{\mathbb{Q}(X = x_i \,|\, Y = 0)} \\ &= \frac{\mathbb{Q}(X > x_{i-1} \,|\, Y = 1) - \mathbb{Q}(X > x_i \,|\, Y = 1)}{\mathbb{Q}(X > x_{i-1} \,|\, Y = 0) - \mathbb{Q}(X > x_i \,|\, Y = 0)} \\ &= \frac{\mathrm{HR}(x_{i-1}) - \mathrm{HR}(x_i)}{\mathrm{FAR}(x_{i-1}) - \mathrm{FAR}(x_i)} \end{aligned}$$

implies that

$$\mathrm{LR}(x_{i+1}) \geq \mathrm{LR}(x_i) \;\Leftrightarrow\; \frac{\mathrm{HR}(x_i) - \mathrm{HR}(x_{i+1})}{\mathrm{FAR}(x_i) - \mathrm{FAR}(x_{i+1}} \geq \frac{\mathrm{HR}(x_{i-1}) - \mathrm{HR}(x_i)}{\mathrm{FAR}(x_{i-1}) - \mathrm{FAR}(x_i)}$$

and the right-hand side is equivalent to the ROC curve being concave, thereby demonstrating the equivalence of (a) and (b). □

Appendix 3.B Equivalence of ROC curve and Murphy diagram dominance for calibrated probability forecasts

Towards the proof of Theorem 3.9, we introduce an alternative characterization of ROC curve dominance.

Lemma 3.19. *For absolutely continuous forecasts, an equivalent characterization of ROC curve dominance is*

$$\mathrm{HR}_1(t) \geq \mathrm{HR}_2(t) + [\mathrm{FAR}_1(t) - \mathrm{FAR}_2(t)] \, \frac{t\,\pi_0}{(1-t)\,\pi_1} \tag{3.15}$$

for all $t \in (0,1)$*. For discrete forecasts it suffices to evaluate (3.15) on the ordered support points* x_i *of forecast* X_1*. This yields*

$$\mathrm{HR}_1(x_i) \geq \mathrm{HR}_2(x_i) + [\mathrm{FAR}_1(x_i) - \mathrm{FAR}_2(x_i)] \, \frac{x_{i+1}\,\pi_0}{(1-x_{i+1})\,\pi_1} \tag{3.16}$$

for all i*, except at any maximum. Here we set* $x_i\,\pi_0/((1-x_i)\,\pi_1) = \infty$ *for* $x_i = 1$*, such that (3.16) implies*

$$\mathbb{Q}(X_1 = 1) \geq \mathbb{Q}(X_2 = 1).$$

Forecast X_1 *strictly dominates* X_2 *if (3.15) or (3.16) holds for all* $t \in (0,1)$ *or all* x_i *and with strict strict inequality for some* $t \in (0,1)$ *or for at least one* x_i*, respectively.*

Proof. As X_1 and X_2 are conditionally calibrated forecasts, their CEPs are nondecreasing functions of t and the corresponding ROC curves R_1 and R_2 are concave. In the following, we treat first the case of absolutely continuous forecasts X_1 and X_2 before considering discrete forecasts.

Fix some $t \in (0,1)$ with corresponding hit and false alarm rate $\mathrm{HR}_2(t)$ and $\mathrm{FAR}_2(t)$. Assume that X_1 dominates X_2 in the ROC curve sense. Which values of $\mathrm{HR}_1(t)$ and $\mathrm{FAR}_1(t)$ are then admissible for R_1? To answer this question, note that the slope $s(t)$ of a ROC curve R corresponding to a conditionally calibrated and absolutely continuous probability forecast is given by

$$s(t) = \frac{\frac{d}{dt}\,\mathrm{HR}(t)}{\frac{d}{dt}\,\mathrm{FAR}(t)} = \frac{f_1(t)}{f_0(t)} = \mathrm{LR}(t) = \frac{t\,\pi_0}{(1-t)\pi_1}, \quad t \in (0,1).$$

As $s(t)$ is decreasing with increasing false alarm rate (or decreasing t), it follows that

$$\mathrm{HR}_1(t) \geq \mathrm{HR}_2(t) + [\mathrm{FAR}_1(t) - \mathrm{FAR}_2(t)] \, \frac{t\,\pi_0}{(1-t)\,\pi_1} \tag{3.15}$$

has to hold. To see this suppose that $\mathrm{HR}_1(t)$ is smaller than the right-hand side in (3.15) and that $\mathrm{FAR}_1(t) < \mathrm{FAR}_2(t)$. Then there exists a $t^* < t$ such that $\mathrm{FAR}_1(t^*) = \mathrm{FAR}_2(t)$ and

$$\begin{aligned}\mathrm{HR}_1(t^*) &\leq \mathrm{HR}_1(t) + [\mathrm{FAR}_1(t^*) - \mathrm{FAR}_1(t)]\,\frac{t\,\pi_0}{(1-t)\,\pi_1} \\ &= \mathrm{HR}_1(t) + [\mathrm{FAR}_2(t) - \mathrm{FAR}_1(t)]\,\frac{t\,\pi_0}{(1-t)\,\pi_1} \\ &< \mathrm{HR}_2(t),\end{aligned}$$

contradicting the assumption that X_1 dominates X_2 in the ROC curve sense. The same argument applies when $\mathrm{FAR}_1(t) \geq \mathrm{FAR}_2(t)$.

If (3.15) holds for all $t \in (0,1)$, this implies that all tangents to the ROC curve R_1 are on or above the ROC curve R_2. It further implies that for all false alarm rates $p \in [0,1]$ the corresponding hit rate of forecast X_1 is greater or equal to the hit rate of forecast X_2, which is equivalent to the definition of ROC curve dominance. An illustration of the equivalent characterization of ROC curve dominance is given in Figure 3.12. The left panel displays the ROC curves corresponding to two continuous and conditionally calibrated probability forecasts and the tangents to the ROC curve of the dominating forecast for selected thresholds.

In the discrete case illustrated in the right panel of Figure 3.12, ROC curves are piecewise linear and all information is contained in the raw ROC diagnostic. Here the slope for any ROC curve segment, including its right end point, is $(x_{i+1}\,\pi_0)/((1-x_{i+1})\,\pi_1)$ for $x_i \leq t \leq x_{i+1}$. Suppose that $\mathbb{Q}(X_2 = 1) = 0$, implying that R_2 has a finite slope everywhere. Then X_1 dominates X_2 in the ROC curve sense if (3.15) holds for all x_i in the support of X_1.

If $\mathbb{Q}(X_2 = 1) > 0$, then R_2 has a vertical line segment from (0,0)' to (0, $\mathbb{Q}(X_2 = 1)$). For X_1 to dominate X_2, (3.16) and $\mathbb{Q}(X_1 = 1) \geq \mathbb{Q}(X_2 = 1)$ need to hold such that R_1 has a greater or equal hit rate for every false alarm rate when compared to R_2. □

Proof of Theorem 3.9. Let $F_{1,0}$ and $F_{1,1}$ as well as $F_{2,0}$ and $F_{2,1}$ denote the conditional distributions of forecasts X_1 and X_2, respectively. The condition of Murphy diagram dominance in (3.5) can be reformulated as

$$\begin{aligned}(3.5) \quad &\Leftrightarrow \quad \theta\pi_0\,[1 - F_{1,0}(\theta) - (1 - F_{2,0}(\theta))] + (1-\theta)\pi_1\,[F_{1,1}(\theta) - F_{2,1}(\theta))] \leq 0 \\ &\Leftrightarrow \quad \theta\,\pi_0\,[F_{2,0}(\theta) - F_{1,0}(\theta)] + (1-\theta)\,\pi_1\,[F_{1,1}(\theta) - F_{2,1}(\theta))] \leq 0 \qquad (3.17)\end{aligned}$$

for every $\theta \in (0,1)$. In the continuous case, X_1 dominates X_2 in the ROC curve sense if (3.15) holds or if for every $t \in (0,1)$

$$\begin{aligned}(3.15) \quad &\Leftrightarrow \quad [1 - F_{1,1}(t) - (1 - F_{2,1}(t))] \geq [1 - F_{1,0}(t) - (1 - F_{2,0}(t))]\,\frac{t\,\pi_0}{(1-t)\pi_1} \\ &\Leftrightarrow \quad [F_{2,1}(t) - F_{1,1}(t)]\,(1-t)\pi_1 \geq [F_{2,0}(t) - F_{1,0}(t)]\,t\,\pi_0 \\ &\Leftrightarrow \quad t\,\pi_0\,[F_{2,0}(t) - F_{1,0}(t)] + (1-t)\,\pi_1\,[F_{1,1}(t) - F_{2,1}(t)] \leq 0.\end{aligned}$$

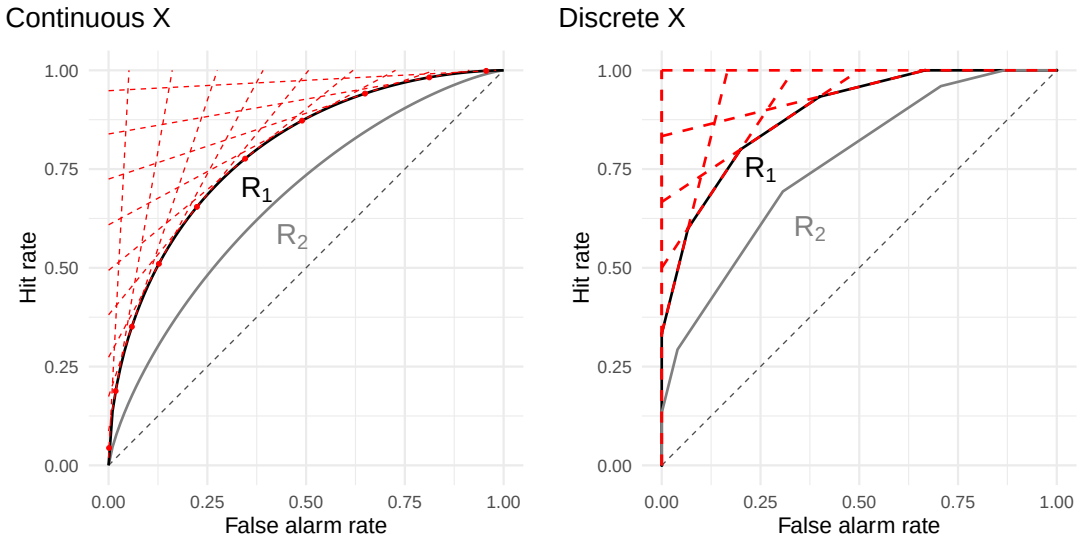


Figure 3.12: Alternative characterization of forecast dominance in the ROC curve sense. Two conditionally calibrated forecasts are depicted in the continuous (left) and the discrete (right) setting. In the discrete setting, all tangents to the dominating ROC curve are displayed, while in the continuous case a subset only is shown.

As this is also the condition for Murphy diagram dominance in (3.17), ROC curve dominance and Murphy dominance are equivalent for conditionally calibrated, absolutely continuous forecasts. For discrete forecasts, this calculation can be restricted to the support points x_i of X_1. As for strict dominance, strict inequality for some $t \in (0, 1)$ in (3.15) is equivalent to strict inequality for $\theta = t$ in (3.17) for the continuous case, and similarly for the discrete case. □

Appendix 3.C Properties of beta ROC curves

Lemma 3.20. *The AUC value for the beta ROC curve is* $\beta/(\alpha + \beta)$.

Proof. We have

$$\mathrm{AUC}(\alpha, \beta) = \int_0^1 B_{\alpha,\beta}(p)\,\mathrm{d}p = \left[pB_{\alpha,\beta}(p) - \frac{\alpha}{\alpha+\beta} B_{\alpha+1,\beta}(p)\right]_0^1 = 1 - \frac{\alpha}{\alpha+\beta},$$

as claimed. □

Lemma 3.21. *The CDF of the beta distribution is concave if* $\alpha \leq 1$ *and* $\beta \geq 2 - \alpha$, *and it is strictly concave if furthermore* $\alpha < 1$.

Proof. The density $b_{\alpha,\beta}$ of the beta distribution satisfies

$$b'_{\alpha,\beta}(x) = \frac{\alpha - 1 + (2 - \alpha - \beta)x}{x(1-x)}\, b_{\alpha,\beta}(x)$$

for $x \in (0, 1)$, from which the statement is immediate. □

Proof of Theorem 3.15. We apply the natural identification and define F_{NI} as in (3.4). Due to the assumption of strong regularity, F_{NI} admits a density on $(0,1)$ that can be extended to a continuous function f_{NI} on $[0,1]$. The arguments in Bernstein's probabilistic proof of the Weierstrass approximation theorem (Levasseur, 1984) show that as $n \to \infty$ the sequence

$$m_n(q) = \frac{1}{n+1} \sum_{k=0}^{n} f_{\mathrm{NI}}\left(\frac{k}{n}\right) b_{k+1,n-k+1}(q)$$

converges to $f_{\mathrm{NI}}(q)$ uniformly in $q \in [0,1]$. Furthermore,

$$a_n = \int_0^1 m_n(q)\,\mathrm{d}q \to \int_0^1 f_{\mathrm{NI}}(q)\,\mathrm{d}q = 1$$

as $n \to \infty$, and for $n = 1, 2, \ldots$ the mapping $p \mapsto M_n(p) = \int_0^p m_n(q)\,\mathrm{d}q/a_n$ represents a mixture of beta CDFs. The uniform convergence of m_n to f_{NI} implies that for every $\epsilon > 0$ there exists an n' such that

$$\begin{aligned}
|F_{\mathrm{NI}}(p) - M_n(p)| &\leq \int_0^p \left| f_{\mathrm{NI}}(q) - \frac{m_n(q)}{a_n} \right| \mathrm{d}q \\
&\leq \int_0^p \left| f_{\mathrm{NI}}(q) - \frac{f_{\mathrm{NI}}(q)}{a_n} \right| \mathrm{d}q + \frac{1}{a_n} \int_0^p |f_{\mathrm{NI}}(q) - m_n(q)|\,\mathrm{d}q \\
&\leq \left| 1 - \frac{1}{a_n} \right| + \frac{1}{a_n} \int_0^p |f_{\mathrm{NI}}(q) - m_n(q)|\,\mathrm{d}q < \epsilon
\end{aligned}$$

for all integers $n > n'$ uniformly in $p \in [0,1]$. The statement of the theorem follows. □

Appendix 3.D Asymptotic normality of minimum distance estimates

Here we demonstrate the asymptotic normality of the minimum distance estimator $\hat{\theta}_n$ in the setting of Subsection 3.3.2. In a nutshell, we apply Theorem 2.2 of Hsieh and Turnbull (1996) and Theorem 3.6 along with the results in Section II in the fundamental paper on minimum distance estimation by Millar (1984). In contrast to the results in Section 4 of Hsieh and Turnbull (1996), which concern minimum distance estimation for the binormal model and ordinal dominance curves, Theorem 3.16 applies to general parametric families and ROC curves.

Proof of Theorem 3.16. We are in the setting of Theorem 2.2 of Hsieh and Turnbull (1996), according to which there exists a probability space with sequences $(B_{1,n})$ and $(B_{2,n})$ of independent versions of Brownian bridges such that

$$\sqrt{n}\,\xi_n(p;\theta_0) = \sqrt{\lambda}\,B_{1,n}(R(p;\theta_0)) + R'(p;\theta_0)\,B_{2,n}(p) + o\big(n^{-1/2}(\log n)^2\big) \quad (3.18)$$

almost surely, and uniformly in p on every interval $[a, b] \subset (0, 1)$. We proceed to verify the regularity conditions for Theorem 3.6 of Millar (1984). As regards the identifiability condition (3.2) and the differentiability condition (3.5) it suffices to note that

$$\xi_n(p;\theta) - \xi_n(p;\theta_0) = R(p;\theta_0) - R(p;\theta)$$

is nonrandom, continuously differentiable with respect to p and the components of the parameter vector θ, and independent of n. The boundedness condition (3.3) is trivially satisfied and the convergence condition (3.4) is implied by (3.18). Finally, we apply[7] (2.17), (2.18), (2.19), and (2.20) in Section II of Millar (1984) to yield (3.10) and (3.11), where the covariance function of the process in (3.9) is

$$\begin{aligned} K(s,t;\theta) &= \operatorname{Cov}(W(s;\theta), W(t;\theta)) \\ &= \lambda \operatorname{Cov}(B_1(R(s;\theta)), B_1(R(t;\theta))) + R'(s;\theta)R'(t;\theta) \operatorname{Cov}(B_2(s), B_2(t)) \\ &= \lambda(\min\{R(s;\theta), R(t;\theta)\} - R(s;\theta)R(t;\theta)) + R'(s;\theta)R'(t;\theta)(\min\{s,t\} - st), \end{aligned}$$

whence $K(s, t; \theta_0)$ is as stated in (3.12). □

The asymptotic result in Corollary 3.17 follows in a straightforward application of the delta method.

[7] We note a typographical error in eq. (2.20) of Millar (1984), where the asymptotic covariance matrix is incorrectly specified as $C^{-1}AC$; it should read $C^{-1}AC^{-1}$ instead.

4 Numerical weather prediction and statistical postprocessing

> Perhaps some day in the dim future it will be possible to advance the computations faster than the weather advances and at a cost less than the saving to mankind due to the information gained. But that is a dream. [1]
>
> Lewis Fry Richardson, 1922

In current practice, weather forecasting relies on ensembles of NWP models. Despite their continuous improvement, systematic errors remain and require statistical postprocessing to realize their full potential. Section 4.1 briefly reviews the principles of NWP ensemble forecasting and discusses The International Grand Global Ensemble (TIGGE) multi-model system and its participating subensembles. Section 4.2 introduces the statistical postprocessing methods Ensemble Model Output Statistics (EMOS) and Bayesian Model Averaging (BMA) and explains the setup of training data composition and parameter estimation as implemented for Chapters 5 and 6. A probabilistic climatological reference forecast is constructed and investigated in Section 4.3.

4.1 Numerical weather prediction and ensembles

Until the beginning of the 20th century, weather forecasts were based on experience in form of weather proverbs or oracles. Bjerknes (1904) introduced the novel idea to describe atmospheric processes by physical laws. Richardson (1922) formulated a numerical model of the atmosphere based on partial differential equations and manually computed a solution by discretizing the atmosphere in space and time. Even though his first prediction was far off and the computation much slower than real time, it can be considered the first NWP forecast. NWP on an operational basis started in the 1950s and since then forecast quality has steadily increased (see, e.g., Figure 1 in Bauer et al., 2015). This improvement was fueled by continuous increases in computational capacity, more observations, new and superior measurement systems, better models and data assimilation, and advances in the understanding of atmospheric processes.

To account for the chaotic nature of the atmosphere in weather forecasting, Leith (1974) proposed a Monte Carlo type approach. Several deterministic NWP

[1] Richardson (1922, p. vi)

model runs are started from slightly different initial conditions, and each prediction represents a potential realization of the future state of the atmosphere. Such a set of deterministic NWP forecasts is called ensemble, and the first operational EPSs were introduced in the 1990s (Toth and Kalnay, 1993; Hamill et al., 2000; Buizza et al., 2000). Nowadays several EPSs use slightly different formulations of the numerical representation of the atmosphere for each deterministic run. This allows to account for the uncertainty in the numerical model formulation. Palmer (2000) reviews fundamental principles of NWP ensembles and Buizza et al. (2005) properties of the EPSs of three leading NWP centers.

An "essentially cost-free approach" to construct an ensemble forecast is to gather the deterministic forecasts of several NWP centers (Ebert, 2001). For the first decade of operational NWP ensemble forecasting these multi-model ensembles were more skillful than the best individual ensemble forecasts (Atger, 1999; Ebert, 2001). The TIGGE multi-model ensemble was set up as part of the THORPEX programme in order to "accelerate improvements in the accuracy of 1-day to 2-week high-impact weather forecasts for the benefit of humanity" (Bougeault et al., 2010, p. 1060). Since its start in October 2006, up to eleven global NWP centers have provided their operational ensemble forecasts, which are accessible on a common $0.5° \times 0.5°$ grid. Park et al. (2008) and Bougeault et al. (2010) discuss objectives and the set-up of TIGGE, including the participating EPSs, in great detail. They also note early results using the TIGGE ensemble, while Swinbank et al. (2016) report on research and achievements accomplished over the last decade. Hagedorn et al. (2012) find that a multi-model ensemble composed of the four best participating TIGGE EPSs, which includes the ECMWF ensemble, outperforms reforecast-calibrated ECMWF forecasts. For the evaluation of NWP precipitation forecast quality as performed in Chapter 5, TIGGE is the most complete and best data source available.

Arguably, the ECMWF EPS is the leading one among the participating TIGGE sub-ensembles (Buizza et al., 2005; Hagedorn et al., 2012; Haiden et al., 2012). It consists of a high-resolution (HRES) run, a control (CNT) run, and 50 perturbed ensemble (ENS) members. The HRES and CNT runs are started from unperturbed initial conditions and differ only in their spatial resolution. The ENS members are started from perturbed initial conditions and have the same spatial resolution as the CNT run. Molteni et al. (1996) and Leutbecher and Palmer (2008) describe generation and properties of the ECMWF EPS in detail.

All other TIGGE sub-ensembles consist only of CNT and ENS forecasts with differing numbers of perturbed ensemble member. Table 4.1 gives an overview over the nine participating TIGGE EPSs that provide accumulated precipitation forecasts and are investigated in Chapter 5.

4.2 Statistical postprocessing

In order to reveal the full potential of ensemble forecasts, we apply statistical postprocessing to raw ensemble forecasts. Statistical postprocessing corrects for

Table 4.1: TIGGE sub-ensembles, with years of availability until 2014, number of ensemble members (number of perturbed members + control run + any high-resolution run), initialization time (UTC), and native grid(s) used in the period 2007–2014.

Source	Acronym	Availability	Members	Init time	Native grid(s)
China Meteorological Administration	CMA	2008–13	14+1	00	TL213/T639
Centro de Previsão Tempo e Estudos Climáticos	CPTEC	2008–14	14+1	00	T126
European Centre for Medium-Range Weather Forecasts	ECMWF	2007–14	50+1+1	00	T399/T639
Japan Meteorological Agency	JMA	2007–13/14	50/26+1	12	TL159/TL319/TL479
Korea Meteorological Administration	KMA	2011–14	16+1	00	N320
Météo France	MF	2010–14	34+1	06	TL798
Meteorological Service of Canada	MSC	2008–14	20+1	00	0.45° uniform
National Centres for Environmental Prediction	NCEP	2008–14	20+1	00	T126
UK Met Office	UKMO	2007–13	23+1	00	N144/N216/N400

systematic miscalibration in form of model biases or incorrect representations of forecast uncertainty and addresses the difference in the spatial scales of model gridboxes and localized observations.

In the following, we review the well established concepts of EMOS (Gneiting et al., 2005) and BMA (Raftery et al., 2005) as well as specific EMOS and BMA models tailored to accumulated precipitation (Scheuerer, 2014; Sloughter et al., 2007). We introduce the BMA and EMOS methods with focus on the 52-member ECMWF EPS and precipitation observations. Adaptations of the postprocessing schemes to other TIGGE sub-ensembles and to the reduced multi-model (RMM) ensemble constructed in Chapter 5 are straightforward, and in case of the BMA model also shortly explained.

We denote the ECMWF HRES, CNT, and ENS members by x_{HRES} or x_{51}, x_{CNT} or x_{52}, and x_{ENS} or $x_1, \ldots, x_{50}$, respectively. We write $\bar{x}_{\mathrm{ENS}}$ for the mean of the ENS members, $\bar{p}$ for the fraction of all 52 members that predict no precipitation, and denote the observed precipitation accumulation by y.

4.2.1 Ensemble Model Output Statistics

EMOS converts an ensemble forecast into a parametric distribution based on the ensemble forecast at hand (Gneiting et al., 2005), and the predictive distribution is of the general form

$$y \,|\, x_{\mathrm{HRES}}, x_{\mathrm{CNT}}, x_{\mathrm{ENS}} \sim g(y \,|\, x_{\mathrm{HRES}}, x_{\mathrm{CNT}}, x_{\mathrm{ENS}}),$$

where the parameters of the predictive density $g(y \,|\, x_{\mathrm{HRES}}, x_{\mathrm{CNT}}, x_{\mathrm{ENS}})$ depend on the ensemble predictions via suitable link functions. For accumulated precipitation forecasts, g must be flexible enough to accommodate a discrete point mass for the probability of no precipitation and a continuous and right-skewed distribution for the amount of precipitation. Scheuerer (2014) introduced the EMOS GEV approach that relies on the three-parameter family of left-censored Generalized Extreme Value (GEV) distributions. The left-censoring allows for a point mass at zero and ensures a non-negative support of the forecast, while the shape parameter allows for flexible skewness. The EMOS GEV model links the mean m and the scale parameter σ of the left-censored GEV distribution to the raw ensemble forecast via

$$m = a_0 + a_{\mathrm{HRES}}\, x_{\mathrm{HRES}} + a_{\mathrm{CNT}}\, x_{\mathrm{CNT}} + a_{\mathrm{ENS}}\, \bar{x}_{\mathrm{ENS}} + a_p\, \bar{p}, \tag{4.1}$$

$$\sigma = b_0 + b_1\, \mathrm{MD}(x_{\mathrm{HRES}}, x_{\mathrm{CNT}}, x_{\mathrm{ENS}}). \tag{4.2}$$

The predictor $\bar{p}$ in (4.1) allows to discriminate between ensemble forecasts where the majority of ensemble members predict very small amounts of precipitation and ensemble forecasts where the majority of members predict no precipitation. The ensemble mean difference (MD)

$$\mathrm{MD}(x_1, \ldots, x_{52}) = \frac{1}{52^2} \sum_{i=1}^{52} \sum_{j=1}^{52} |x_i - x_j| \tag{4.3}$$

is more robust than the standard deviation, though still sensitive to all ensemble members. The shape parameter ξ is not linked to the ensemble forecast, but estimated from training data as explained in Subsection 4.2.4.

For illustration, Fig. 4.1a shows an EMOS GEV postprocessed forecast distribution for 5-day accumulated precipitation at Ouagadougou, Burkina Faso. The 52 raw ECMWF ensemble members are represented by blue marks; they include eleven values in excess of 200 mm, with the CNT member being close to 500 mm. The ensemble forecast at hand informs the statistical parameters of the EMOS postprocessed forecast distribution, which includes a tiny point mass at zero, and a censored GEV density for positive precipitation accumulations, with the 90th percentile being at 174 mm.

4.2.2 Bayesian Model Averaging

A BMA predictive distribution is a weighted sum of component distributions, each of which depends on a single ensemble member (Raftery et al., 2005). For the ECMWF ensemble, this corresponds to

$$\begin{aligned} y \,|\, x_{\text{HRES}}, x_{\text{CNT}}, x_{\text{ENS}} \sim{}& w_{\text{HRES}}\, g_{\text{HRES}}(y \,|\, x_{\text{HRES}}) \\ &+\; w_{\text{CNT}}\, g_{\text{CNT}}(y \,|\, x_{\text{CNT}}) + \frac{w_{\text{ENS}}}{50} \sum_{i=1}^{50} g_{\text{ENS}}(y \,|\, x_i), \end{aligned} \tag{4.4}$$

with nonnegative weights w_{HRES}, w_{CNT}, and w_{ENS} that sum to 1, and reflect the members' performance in the training period. For accumulated precipitation forecasts Sloughter et al. (2007) proposed a BMA Gamma0 model where each of the component distributions consists of a point mass at zero and a Gamma distribution that specifies positive accumulation amounts. The probability of no precipitation $p_k = \mathbb{P}(y = 0 \,|\, x_k)$ for $k \in \{\text{HRES, CNT, ENS}\}$ is estimated independently for each member forecast x_k by logistic regression

$$\text{logit}\; \mathbb{P}(y = 0 \,|\, x_k) = a_{0k} + a_{1k}\, x_k^{1/3} + a_{2k}\, \mathbb{1}(x_k = 0).$$

Sloughter et al. (2007) find the cube-root transform of ensemble member forecast x_k to be a better predictor than other power-transformations including the identity. The indicator function of member forecast x_k being zero allows for a better discrimination between forecast x_k predicting no or small precipitation amounts.

The specification for positive accumulation amounts is based on a Gamma density for the cube-root transformed precipitation amount $\tilde{y} = y^{1/3}$. The Gamma density with shape parameter α_k and scale parameter β_k is

$$h_k(\tilde{y}) = \frac{\tilde{y}^{\alpha_k - 1} \exp(-\tilde{y}/\beta_k)}{\beta_k^{\alpha_k} \Gamma(\alpha_k)} \tag{4.5}$$

for $\tilde{y} > 0$ and 0 else. Its mean $\mu_k = \alpha_k \beta_k$ and variance $\sigma_k^2 = \alpha_k \beta_k^2$ are a linear function of the (cube-root transformed) member forecast x_k given by

$$\begin{aligned} \mu_k &= b_{0k} + b_{1k} x_k^{1/3}, \\ \sigma_k^2 &= c_0 + c_1 x_k. \end{aligned}$$

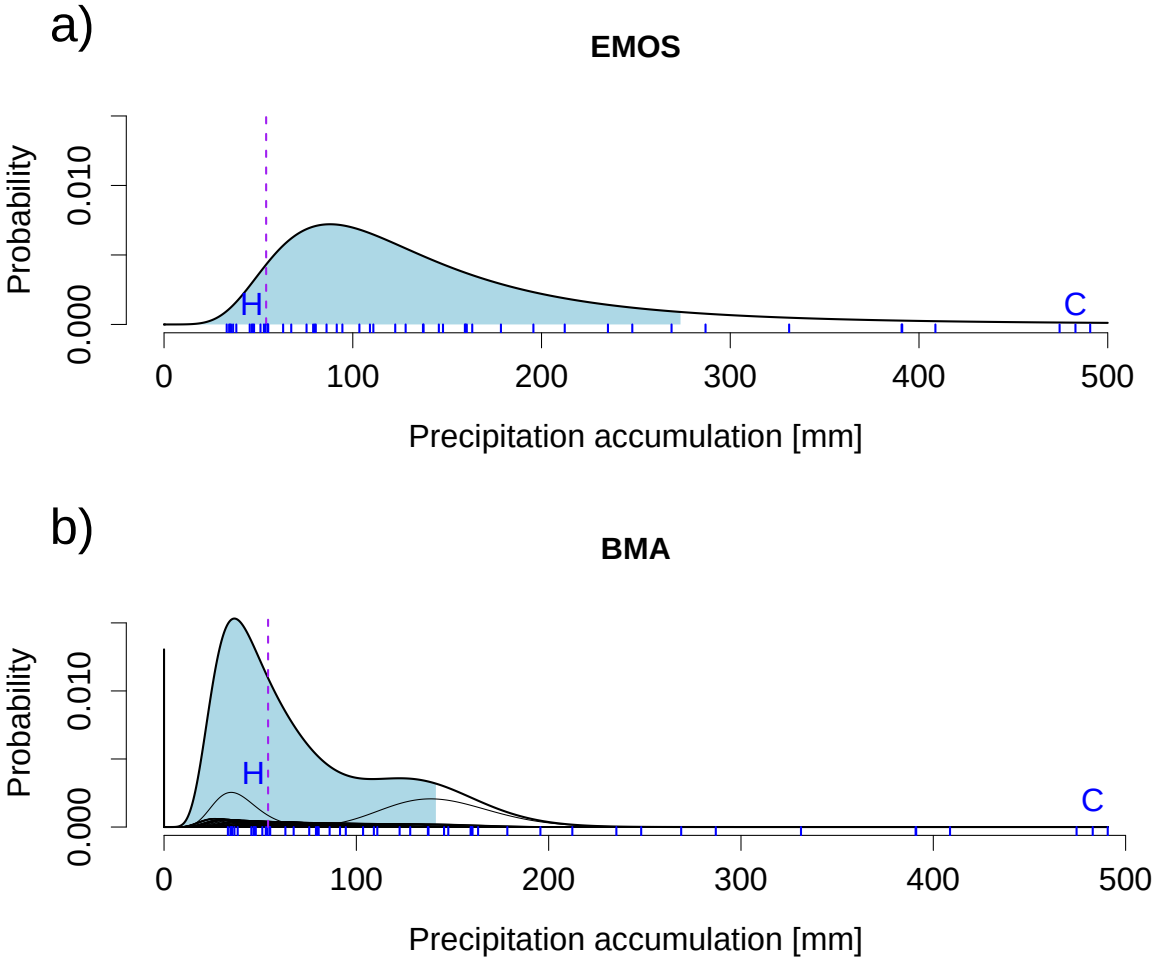


Figure 4.1: EMOS and BMA postprocessed ECMWF ensemble forecasts for 5-day accumulated precipitation at Ouagadougou, Burkina Faso, valid 03 Aug – 08 Aug 2007. The blue ticks at the bottom represent the 52 raw ECMWF ensemble members, including the HRES (H) run, the CNT (C) run, and the 50 perturbed ENS members. (a) The EMOS postprocessed forecast includes a tiny point mass at zero and a censored GEV density for positive accumulations. (b) The BMA postprocessed forecast includes a point mass at zero, which is represented by the solid bar, and a mixture of power transformed Gamma densities for positive accumulations. The 52 component densities are represented by the thin black curves, with the HRES and CNT components standing out. The lower 90% prediction interval is indicated in light blue, and the dashed bar represents the verifying precipitation accumulation.

While the statistical coefficients for the mean μ_k of each Gamma model are estimated for g_{HRES}, g_{CNT}, and g_{ENS} separately, the coefficients for the variance of the Gamma model are shared. The discrete-continuous distribution of each component g_k for the cube-root transformed accumulation amount $\tilde{y}$ is given by

$$g_k(\tilde{y} \,|\, x_k) = p_k \, \mathbb{1}(\tilde{y} = 0) + (1 - p_k) \, \mathbb{1}(\tilde{y} > 0) \, h_k(\tilde{y}) \tag{4.6}$$

To obtain the BMA predictive distribution for the precipitation accumulation in the unit mm, rather than the cube root thereof, a backtransformation is applied as described by Sloughter et al. (2007).

Figure 4.1b shows such a BMA postprocessed forecast distribution for the aforementioned forecast case at Ouagadougou, Burkina Faso. The postprocessed distribution involves a point mass of about 0.01 at zero, and a mixture of power transformed Gamma densities for positive accumulations, with the 90th percentile being at 141 mm. In this example, the BMA and EMOS postprocessed distributions are sharper than the raw ECMWF ensemble, and nevertheless the verifying observation is well captured.

Adaptations to other ensembles considered in Chapter 5 are straightforward as described by Fraley et al. (2010). For example, in the case of the RMM ensemble with 15 different members, each member receives its own component distribution, BMA weight, logistic regression coefficients for the probability of no precipitation, and statistical parameters for the Gamma mean model, whereas the coefficients for the Gamma variance model are shared.

4.2.3 Training data

Statistical postprocessing of raw ensemble forecasts by EMOS or BMA necessitates the estimation of statistical parameters on a set of training data. Several approaches exist for the composition of training data and we distinguish local and regional ones. For the local approach, training data contain only forecast–observation-pairs from the considered location, while they stem from all locations in the considered region for the regional approach. To account for temporal changes of raw ensemble forecast errors, a rolling training period is employed, where the training data consist of the n most recent days for which data are available at initialization time.

In Chapter 5, we employ the regional approach with a rolling training period of $n = 20$ days. The choice of 20 days is consistent with the literature (e.g., Thorarinsdottir and Gneiting, 2010) and we assess its appropriateness for EMOS and BMA postprocessed precipitation forecasts for northern tropical Africa. We rely for verification on station and satellite-based $0.25° \times 0.25°$ Tropical Rainfall Measurement Mission (TRMM) observations (see Section 5.2 for details) and accumulation periods of 1 and 5 days. Results are displayed separately for West Sahel, East Sahel, and Guinea Coast. For the geographic location of the stations within and the spatial extent of these three regions, see Figure 5.1.

Figure 4.2 displays BS skill of EMOS postprocessed ECMWF ensemble forecasts for the occurrence of precipitation with rolling training periods of 10 to 50

days in increments of 5 days relative to the same forecast with $n = 20$ days. Independent of the accumulation period, we use here and in the following a threshold of 0.2 mm to determine the occurrence of precipitation and find only minimal changes under other choices of the threshold between 0.0 mm and 1.0 mm. In all panels, BS skill varies from monsoon season to monsoon season around the level of neutral skill. Consequently, the average BS skill across 2007–2014 (black line) is typically close to zero and reveals no systematic improvement in BS skill for any of the examined training periods. Training periods of less than 20 days often deteriorate predictive performance. Figure 4.3 displays CRPS skill of EMOS GEV postprocessed forecasts for the amount of precipitation in the same setting as Figure 4.2. Again, CRPS skill varies from monsoon season to monsoon season, but training periods of less than 20 days systematically deteriorate postprocessed forecast skill.

Figure 4.4 displays BS skill of BMA Gamma0 postprocessed predictions in the same setting as Figure 4.2. As BMA Gamma0 is more computationally intense than EMOS GEV, training periods are restricted to 20, 30, and 40 days. Interestingly, the interannual variability in BS skill of BMA PoP forecasts is lower than for EMOS PoP forecasts. Training periods of 30 and 40 days yield no systematic improvement or deterioration of BS skill. CRPS skill of BMA postprocessed forecasts as displayed in Figure 4.5 only confirms previous findings. In summary, these results show that our findings in Chapter 5 are quite insensitive to the choice of n when using training periods between 20 and 50 days, while for training periods shorter than $n = 20$ days a clear deterioration can be observed. The advantage of the regional over local approach is that it requires much shorter training periods. Experiments for local postprocessing in the setting of Chapter 5 (not shown here) yield very similar results as regional postprocessing.

In Chapter 6, we rely on a semi-local approach for the composition of training data in the tropics. Out of the surrounding eight $1^\circ \times 1^\circ$ gridboxes, we consider only those that belong to the same surface type (land, ocean) as the central gridbox and use the $n = 500$ most recent forecast–observation-pairs available at initialization time from each of those gridboxes.

4.2.4 Parameter estimation

For the estimation of the EMOS GEV model parameters, we rely on CRPS minimization. A closed form expression of the CRPS is due to Scheuerer (2014) and allows for efficient computation. For BMA Gamma0, the parameters a_{0k}, a_{1k}, a_{2k} as well as b_{0k} and b_{1k} are estimated separately for each ensemble member by logistic and linear regression, respectively. As the variance parameters vary typically only little across forecast members, all ensemble members share the same c_0 and c_1. For the estimation of the weights $w_{\text{HRES}}, w_{\text{CNT}}, w_{\text{ENS}}$ as well as the variance parameters c_0 and c_1, we rely on maximum likelihood, implemented via the expectation-maximization (EM) algorithm developed by Sloughter et al. (2007).

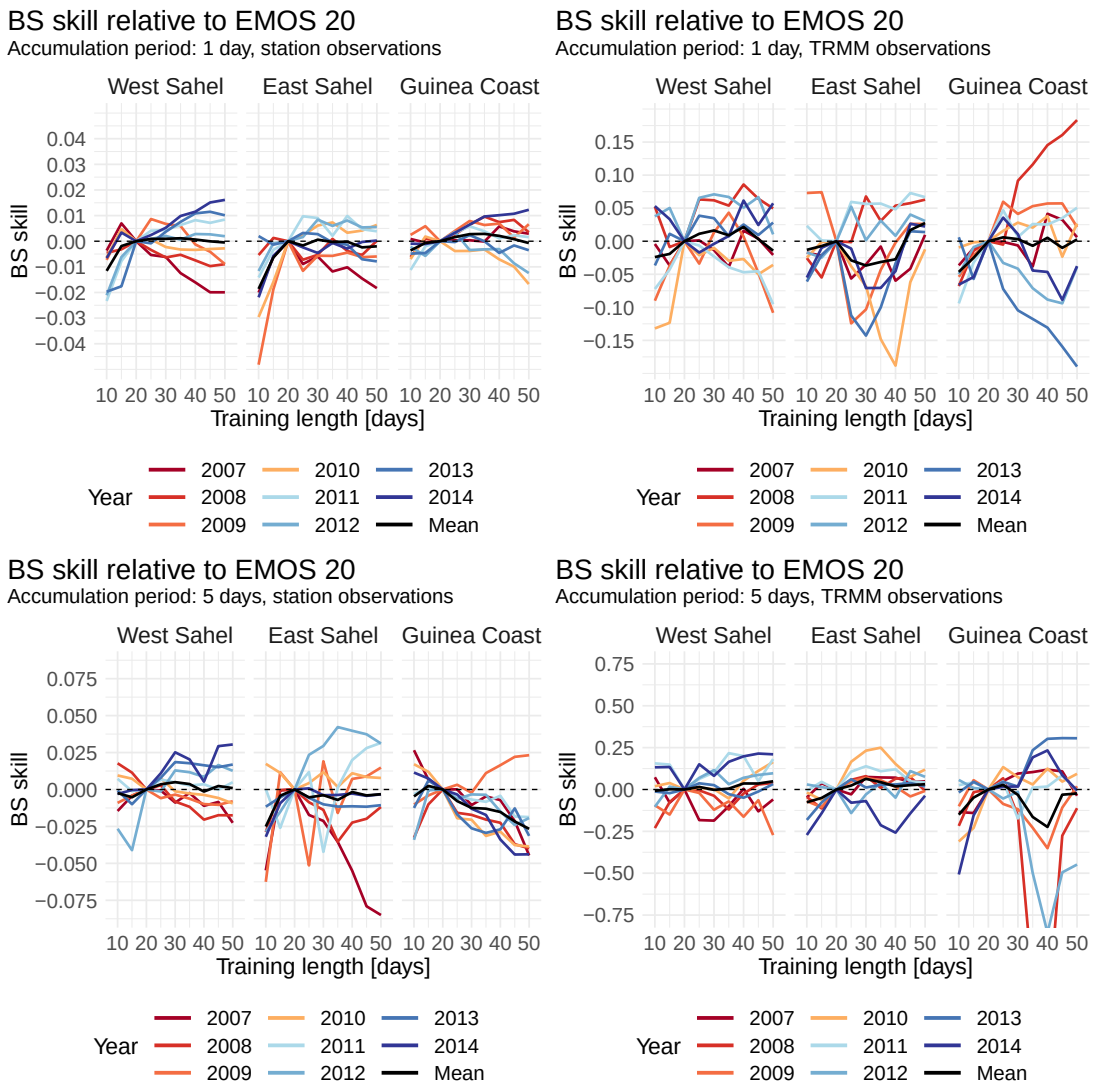


Figure 4.2: BS skill of EMOS GEV postprocessed ECMWF ensemble forecasts for the occurrence of precipitation with rolling training periods of 10 to 50 days in increments of 5 days relative to the same forecast with $n = 20$ days. Results are stratified by region and year, verified against station (left) and $0.25° \times 0.25°$ TRMM (right) observations for accumulation periods of one (top) and five (bottom) days. Details on station and TRMM observations are provided in Section 5.2 and the geographic location and spatial extent of West Sahel, East Sahel, and Guinea Coast is displayed in Figure 5.1.

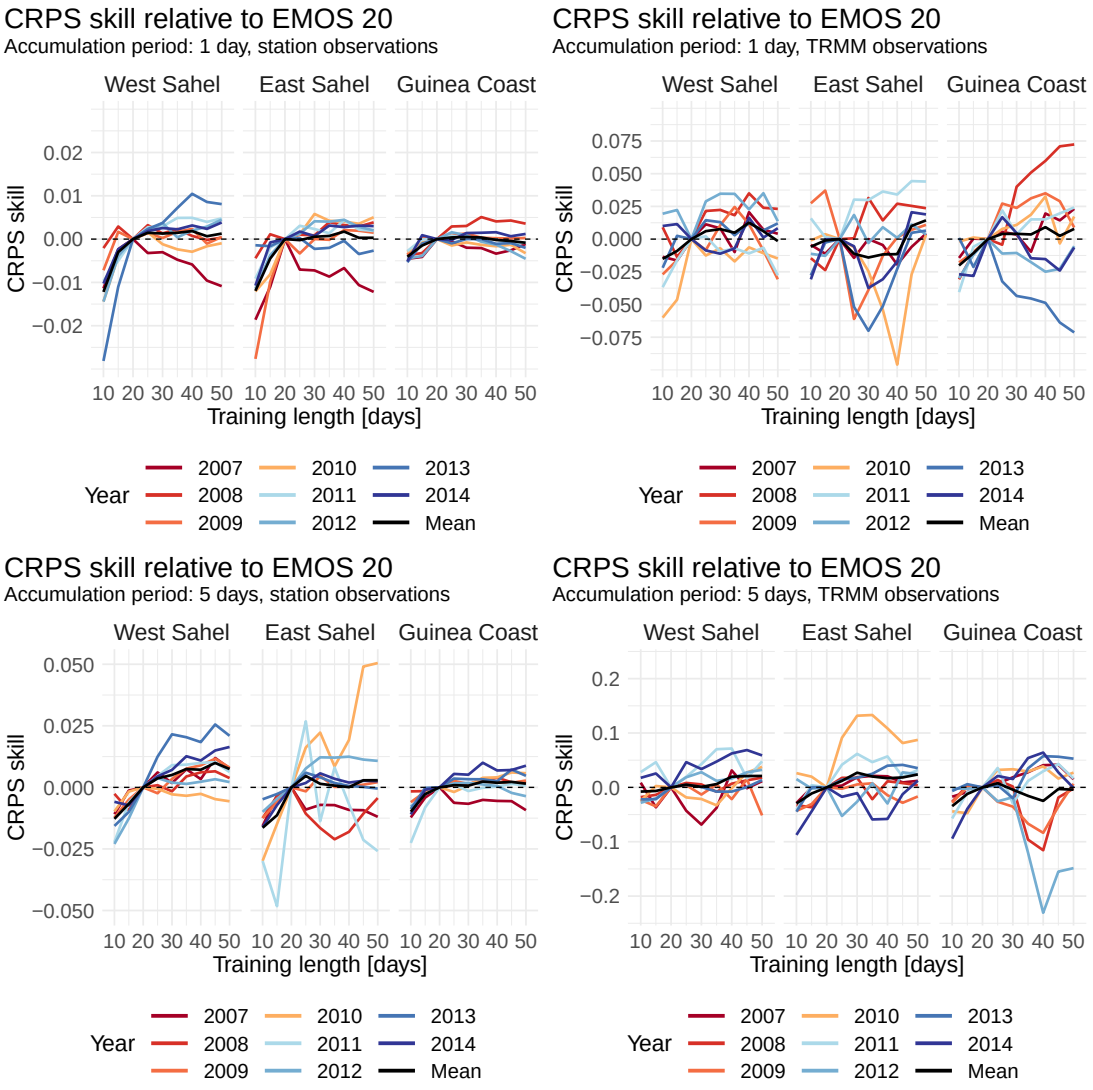


Figure 4.3: Same as Figure 4.2, but for CRPS skill of EMOS GEV postprocessed forecasts for the amount of precipitation. © Copyright 2018 AMS.

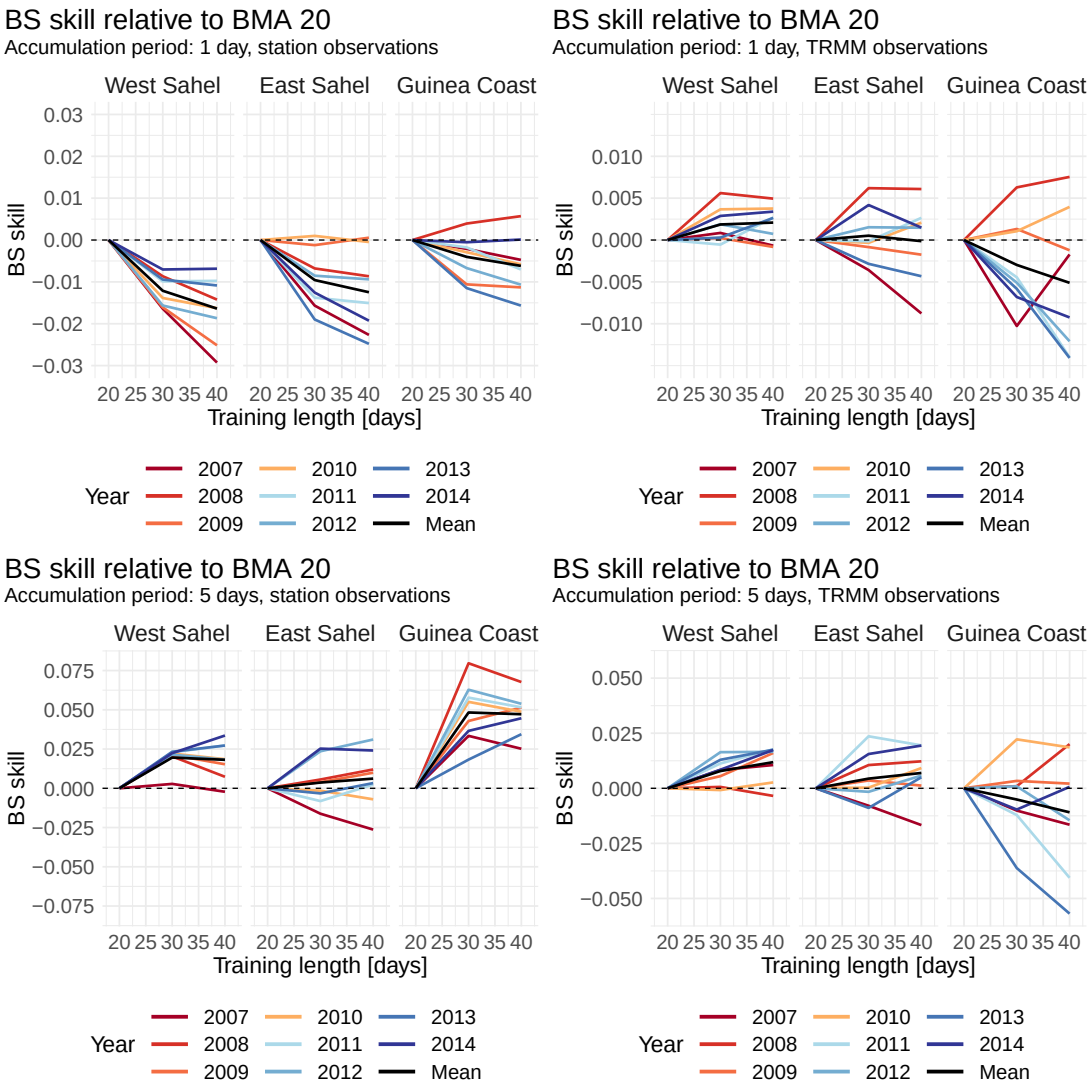


Figure 4.4: BS skill of BMA Gamma0 postprocessed ECMWF ensemble forecasts for the occurrence of precipitation with rolling training periods of $n \in \{20, 30, 40\}$ days relative to the same forecast with $n = 20$ days. Results are stratified by region and year, verified against station (left) and $0.25° \times 0.25°$ TRMM (right) observations for accumulation periods of one (top) and five (bottom) days. © Copyright 2018 AMS.

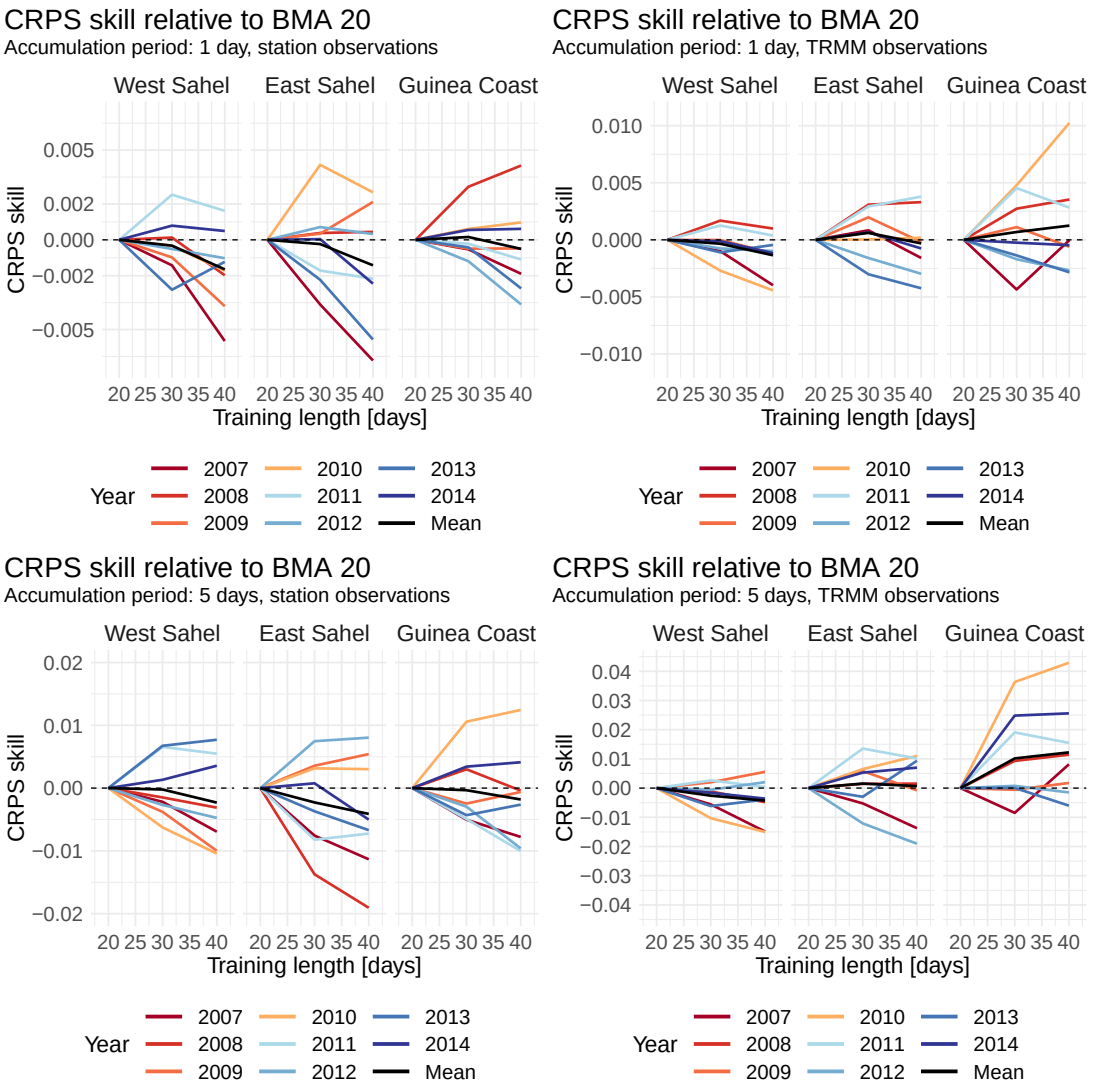


Figure 4.5: Same as Figure 4.4, but for CRPS skill of BMA Gamma0 postprocessed forecasts for the amount of precipitation. © Copyright 2018 AMS.

4.3 Probabilistic climatological reference forecast

For the assessment of raw and postprocessed ensemble forecast skill, the availability of a good benchmark forecast is essential. In the following, we introduce the concept of the Extended Probabilistic Climatology (EPC) forecast and discuss its properties for the specific implementations in Chapters 5 and 6.

Consider a probabilistic climatology that consists of the observations during the 30 years prior to the considered year at the considered day of the year and the considered station. It can be understood as a 30-member observation-based ensemble forecast that represents the climatological distribution of rainfall at a given location and date, but does not incorporate dynamic information about the state of the atmosphere. We extend this (standard) probabilistic climatology by including observations in a $\pm$2-day window around the considered day for the 30 years prior to the considered year, and refer to this as EPC.

In Chapter 5, we rely on station observations and satellite-based TRMM observations for the assessment of NWP ensemble forecast quality for three regions in northern tropical Africa.[2] Hamill and Juras (2006) note that pooling can lead to a deterioration when performed across data with differing climatologies, leading to a perceived, but incorrect improvement of assessed model forecast skill. In case of the EPC, however, neighboring daily climatologies can be assumed to be very similar and the pooling is performed over a range of $\pm$2 days only.

To assess the correctness of this assumption, we evaluate the skill of EPC forecasts with window lengths between 0 and $\pm$20 days for 1- and 5-day accumulated precipitation forecasts relative to the proposed EPC forecast with a window length of $\pm$2 days. Figure 4.6 displays the results for the EPC forecast based on and verified against station observations. For an accumulation period of one day, BS and CRPS skill is negative for a standard probabilistic climatology in all regions, and for most years and regions positive for EPC with window lengths of more than $\pm$2 days. Mean CRPS and BS skill across 2007–2014 (black line) displays a better performance of EPC forecasts with window lengths of more than $\pm$2 days, but improvement is lower than 1%. For 5-day accumulations, standard probabilistic climatologies underperform relative to the proposed EPC forecast. Window lengths of more than $\pm$2 days are beneficial for most years, but average BS and CRPS skill across 2007-2014 is lower than 2% in all regions. While a slightly larger window would improve the skill of the reference forecast for most regions, leadtimes, and years, average improvement is small and hence our conclusions in Chapter 5 are quite insensitive to this choice. As TRMM observations are available for the period 1998–2014 only, the TRMM-based EPC relies on this period but without the considered verification year. The assessment of $\pm$2 days being a reasonable choice remains.

For the investigation in Chapter 6, TRMM observations are available for the period 1998–2017. We construct TRMM-based EPC forecasts with window lengths of 0 to $\pm$40 days in increments of $\pm$5 days for each pixel as before, but rely on

[2] See Figure 5.1 for the geographic location and spatial extent of the three regions.

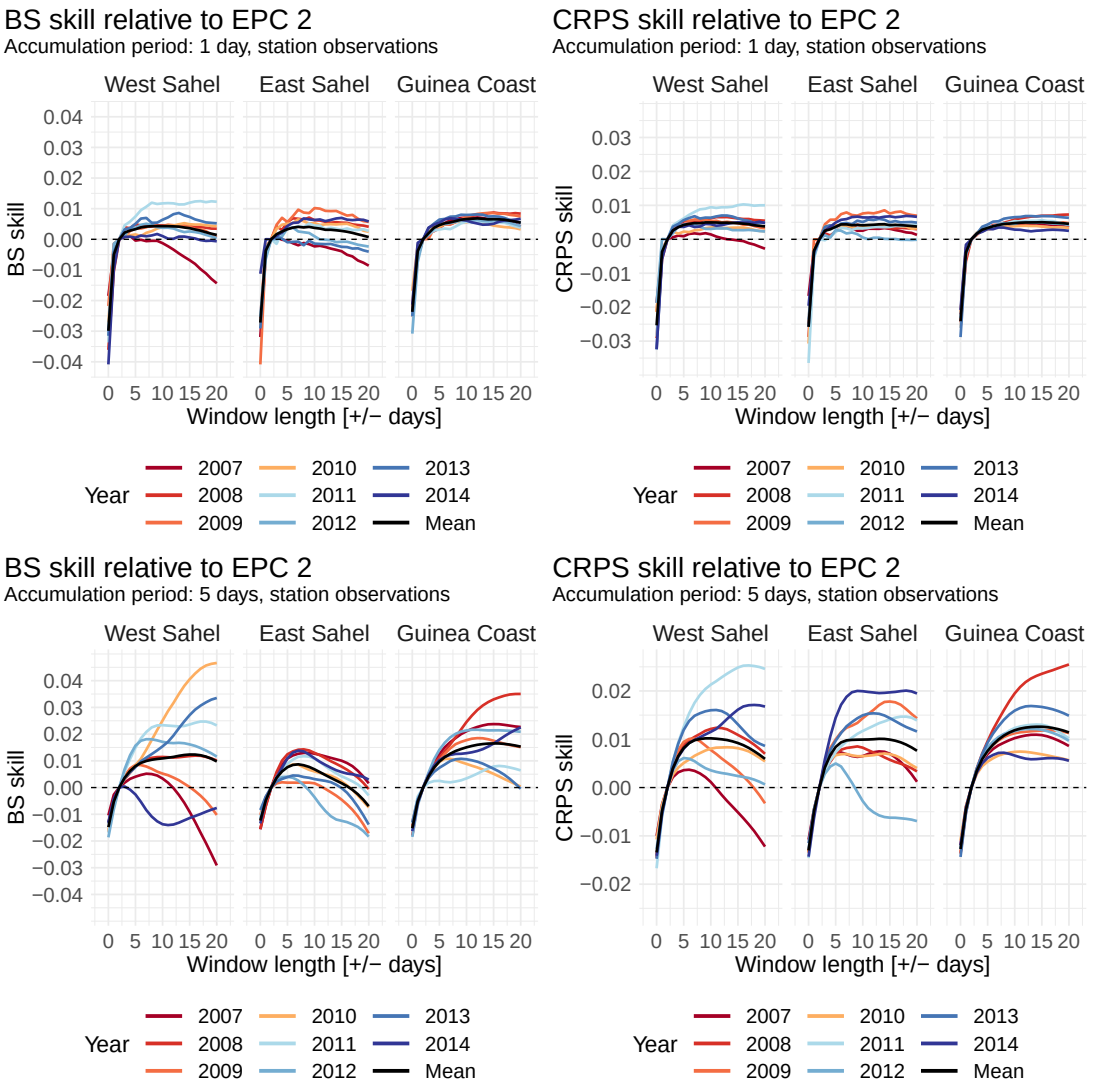


Figure 4.6: BS and CRPS skill for EPC forecasts with window lengths of 0 to ±20 days relative to the proposed EPC forecast. Results are stratified by region and year and displayed for accumulation periods of one (top) and five (bottom) days. Forecasts are issued for and verified against station observations. © Copyright 2018 AMS.

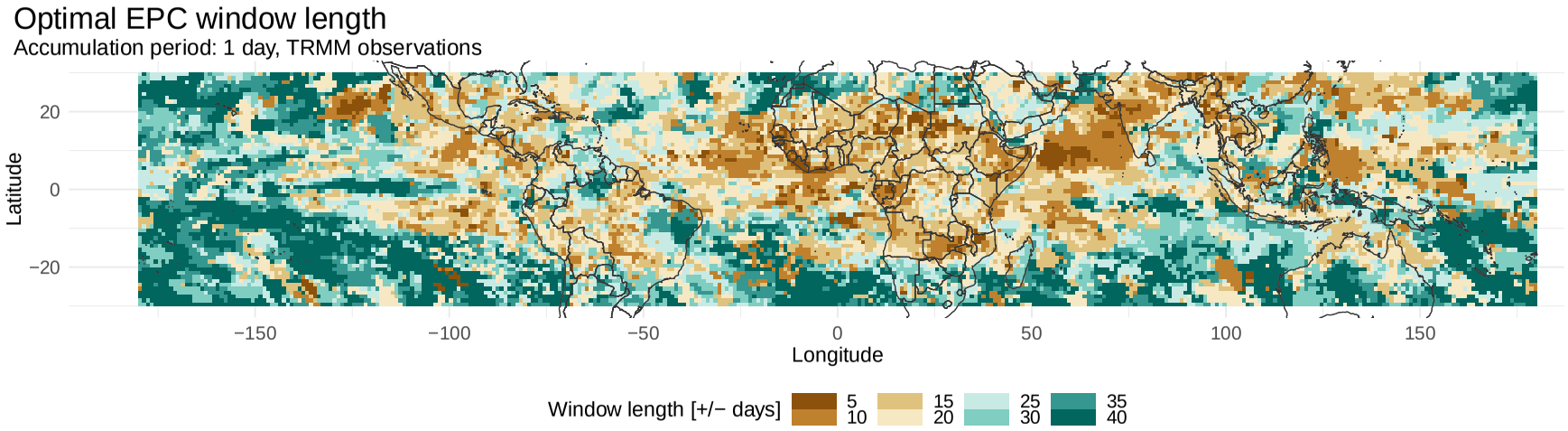


Figure 4.7: Optimal window length of the TRMM-based EPC forecast. Displayed is the optimal window length in days of the TRMM-based EPC forecast for 1-day accumulated precipitation as evaluated by the best (lowest) mean CRPS for 1998–2017.

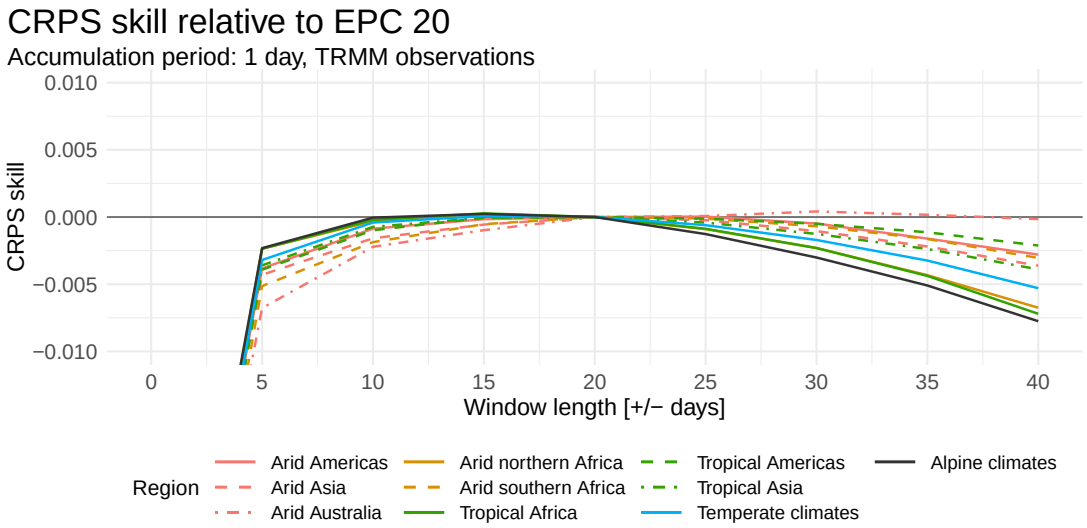


Figure 4.8: CRPS skill for TRMM-based EPC forecasts with window length of $0, \pm 5, \ldots, \pm 40$ days relative to EPC forecasts with ± 20 days. Results are stratified by Köppen-Geiger climates introduced in Chapter 6.

1998–2017 instead of 1998–2014. As reference forecast in Chapter 6, we use for each pixel the EPC forecast with the window length that achieved the best CRPS mean score. Figure 4.7 displays the optimal window lengths for TRMM-based EPC forecasts throughout the tropics. Except for a window length of 0 days, all examined window lengths turn out to be optimal for several gridboxes. In general, window lengths of ± 5 to ± 20 days are preferable for tropical climates with a strong seasonal evolution of rainfall, while window lengths of more than ± 20 days are beneficial for climates with no or only a weak seasonality of rainfall. Over ocean, the spatial distribution is less clear and longer window lengths turn out to be optimal for the central Pacific ocean, while shorter ones are beneficial over the tropical Atlantic and parts of the Indian ocean. For details on the definition of climatic regions, see Subsection 6.2.4 and Figure 6.2. Figure 4.8 displays for each of the climatic regions the CRPS skill of EPC forecasts with a window length of 0 to ± 40 days in increments of ± 5 days relative to an EPC forecast with ± 20 days. Except for standard probabilistic climatologies, average skill of EPC forecasts in each Köppen-Geiger climate is within $\pm 1\%$ throughout.

5 Skill of global raw and postprocessed ensemble predictions of rainfall over northern tropical Africa

In this chapter, we investigate the skill of global raw and postprocessed ensemble predictions of rainfall over northern tropical Africa. For the agriculturally dominated societies in this region, accumulated precipitation forecasts potentially have high socioeconomic benefit. We analyze the performance of nine operational global EPSs relative to climatology-based forecasts for 1 to 5-day accumulated precipitation based on the monsoon seasons 2007–2014 for three regions within northern tropical Africa. To assess the full potential of raw ensemble forecasts across spatial scales, we apply state-of-the-art statistical postprocessing methods in form of BMA and EMOS, and verify against station and spatially aggregated, satellite-based gridded observations.

5.1 Introduction

The bulk of precipitation in the tropics is related to moist convection, in contrast to the frontal-dominated extratropics. Due to the small-scale processes involved in the triggering and growth of convective systems, quantitative precipitation forecasts are known to have overall poorer skills in tropical latitudes (Haiden et al., 2012). This can be monitored in quasi-real time on the World Meteorological Organization (WMO) Lead Centre on Verification of Ensemble Prediction System website (`http://epsv.kishou.go.jp/EPSv`) by comparing deterministic and probabilistic skill scores for 24-hour precipitation forecasts for the 20°N–20°S tropical belt with those for the northern and southern hemisphere extratropics. There are hints that precipitation and cloudiness forecasts in the tropics show enhanced skill during regimes of stronger synoptic-scale forcing (Söhne et al., 2008; Davis et al., 2013; van der Linden et al., 2017) or in regions of orographic forcing (Lafore et al., 2017), but large parts of the tropical land masses are dominated by convection that initiates from small-scale surface and boundary layer processes and sometimes organizes into mesoscale convective systems (MCSs). The latter depends mostly on the thermodynamic profile and vertical wind shear (Maranan et al., 2018).

In this context, northern tropical Africa, particularly the semi-arid Sahel, can

be considered a region where precipitation forecasting is particularly challenging. The area consists of vast flatlands, MCSs during boreal summer provide the bulk of the annual rainfall (Mathon et al., 2002; Fink et al., 2006; Houze et al., 2015), and convergence lines in the boundary layer or soil moisture gradients at the km-scale can act as triggers for MCSs (Lafore et al., 2017). Sahelian MCSs often take the form of meridionally elongated squall lines with sharp leading edges characterized by heavy rainfall. Synoptic-scale African easterly waves (AEWs) are known to be linked to squall line occurrence in the western Sahel (Fink and Reiner, 2003) and lead to an enhanced skill of cloudiness forecasts over West Africa (Söhne et al., 2008).

However, NWP models are known to have an overall poor ability to predict rainfall systems over northern Africa. For example, the gain in skill by improved initial conditions due to an enhanced upper-air observational network during the 2006 African Monsoon Multidisplinary Analysis (AMMA) campaign (Parker et al., 2008) was lost in NWP models after 24 hours of forecast time, potentially due to the models' inability to predict the genesis and evolution of convective systems (Fink et al., 2011). Given the substantial challenges involved in forecasting rainfall in northern Africa, one might hope that EPSs provide an accurate assessment of uncertainties and a more useful forecast overall. Despite many advances in the generation of EPSs, ensembles share structural deficiencies and require statistical postprocessing to realize the full potential of ensemble forecasts (Gneiting and Raftery, 2005). Additionally, statistical postprocessing performs implicit downscaling from the model grid resolution to finer resolutions or station locations. In the following, we will explore whether established methods such as BMA and EMOS can improve precipitation forecasts for northern tropical Africa. To our knowledge, the investigation presented here and in Vogel et al. (2018) is the first study to rigorously and systematically assess the quality of ensemble forecasts for precipitation over northern tropical Africa. This is partly related to the fact that for this region ground verification data from rain gauge observations are infrequent on the Global Telecommunication System (GTS), the standard verification data source for NWP centers.

The ultimate goal of this chapter is to provide an exhaustive assessment of our current ability to predict rainfall over northern tropical Africa, considering the skill of raw and postprocessed forecasts from TIGGE. Any skill, if existing, would be expected to come from resolved large-scale forcing processes as mentioned above. We examine accumulation periods of 1- to 5-days for the monsoon seasons 2007–2014 and verify against about 21,000 daily rainfall observations from 132 rain gauge stations and satellite-based gridded precipitation observations. Section 5.2 introduces the RMM ensemble forecast based on the TIGGE ensemble as well as station and satellite-based observations used for verification. For postprocessing, we rely on EMOS and BMA and compare against EPC as our benchmark forecast. These methods are explained in detail in Chapter 4. Their verification is based on proper scoring rules, consistent scoring functions, uPIT histograms, and reliability, Murphy, and ROC diagrams introduced in Chapter 2. Results are presented in Section 5.3, where we verify 1-day accumulated ECMWF

precipitation forecasts against station observations. This analysis is performed in particular depth and serves as a fundamental exemplar. We also evaluate ECWMF ensemble forecasts at longer accumulation times and for spatial aggregations, before turning to the analysis of all TIGGE sub-ensembles. Implications of our findings and possible alternative methods for forecasting precipitation over northern tropical Africa are discussed in Section 5.4.

5.2 Data

5.2.1 Forecasts

Of the eleven participating ensembles of TIGGE, nine provide accumulated precipitation forecasts (see Table 4.1). In addition to the separate evaluation of each participating TIGGE sub-ensemble, we rely on the RMM ensemble to evaluate the benefit from intermodel variability. For each of the seven sub-ensembles available for the period 2008–2013, the RMM ensemble uses the mean of the perturbed members, and the control run, and in case of the ECMWF EPS furthermore the high-resolution run, as individual contributors. The RMM ensemble therefore consists of 15 members and, as postprocessing performs an implicit weighting of all contributions, a manual selection of sub-ensembles as performed by Hagedorn et al. (2012) is not necessary.

5.2.2 Observations

Despite multiple advances in satellite rainfall estimation, station observations of accumulated precipitation remain a reliable and necessary source of information. However, the meteorological station network in tropical Africa is sparse and clustered, and often observations of many stations are not distributed through the GTS. The Karlsruhe African Surface Station Database (KASS-D) contains precipitation observations from a variety of networks and sources. Manned stations operated by African national weather services provide the bulk of the 24-hour precipitation data. Due to long-standing collaborations with these services and African researchers, KASS-D contains many observations not available in standard, GTS-fed station databases. Within KASS-D, 960 stations have daily accumulated precipitation observations and usually these are measured between 06–06 universal time coordinated (UTC).

After excluding stations outside the study domain, and removing sites with less than 80% available observations in any of the monsoon seasons, the remaining 132 stations were subject to quality control, as described in the Appendix 5.A, and passed these tests. Based on their rainfall climate (e.g. Fink et al., 2017) and geographic clustering, the stations were assigned to three regions, as indicated in Figure 5.1: West Sahel, East Sahel, and Guinea Coast.

As NWP forecasts are issued for grid cells, the comparison of station observations against gridded forecasts is fraud with problems. To allow for an additional assessment of forecast quality without a gauge-to-gridbox comparison and for

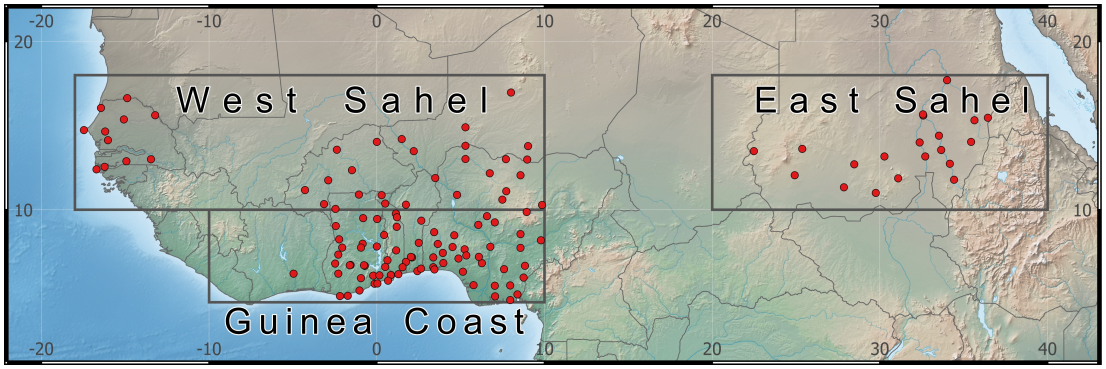


Figure 5.1: Geographical overview of the study domain, with the locations of the observation stations (•) within the three considered regions. © Copyright 2018 AMS.

areas without station observations, we use satellite-based, gridded precipitation estimates. Based on recent studies, version 7 (and also version 6) of the TRMM 3B42 gridded data set is regarded the best available satellite precipitation product for northern tropical Africa, despite a small dry bias (Roca et al., 2010; Maggioni et al., 2016; Engel et al., 2017).

TRMM merges active measurements from the precipitation radar with passive, radar-calibrated information from infrared as well as microwave measurements (Huffman et al., 2007). Based on monthly accumulation sums, TRMM estimates are calibrated against nearby gauge observations. TRMM 3B42-V7 data are available on a $0.25° \times 0.25°$ grid with three hourly temporal resolution.

5.2.3 Data preprocessing

Based on 1-day accumulated station observations, we derive 2- to 5-day accumulated precipitation observations by summing over consecutive 1-day observations. As these cover the period from 06 UTC of the previous day to 06 UTC of the considered day and as all TIGGE sub-ensembles, except Météo France (MF), have initialization times different from 06 UTC, we use the most recent run available at that time, and adapt accordingly. Specifically, for the sub-ensembles initialized at 00 UTC, we use the difference between the 30-hour accumulated and the 6-hour accumulated precipitation forecast. For initialization at 12 UTC, we use the difference between the 42-hour accumulated and the 18-hour accumulated precipitation forecast, and for longer accumulation times, we extend correspondingly.

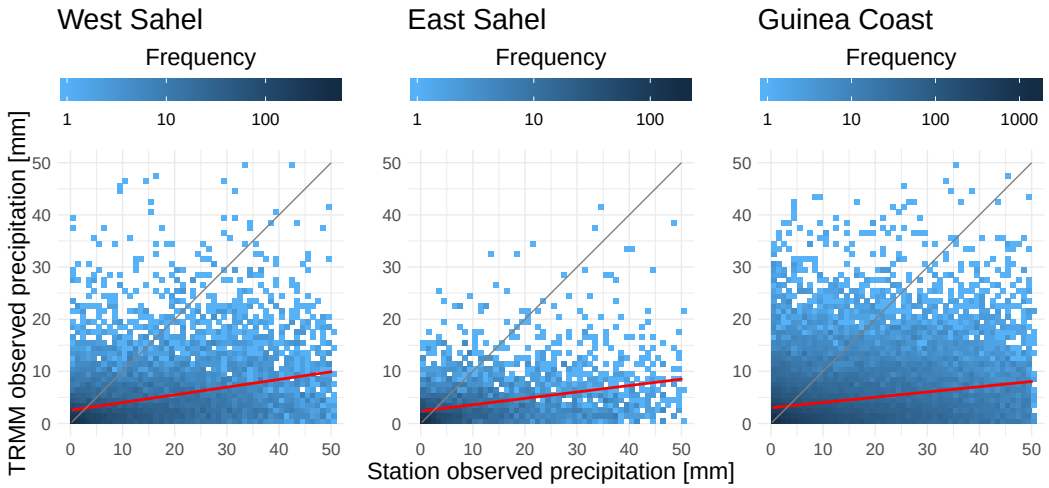
To obtain forecasts for a specific station location from gridded NWP forecasts, bilinear interpolation as well as a nearest neighbor approach are possible. We use the latter, implying that the forecast for the station is the same as the forecast for the grid cell containing the station. Especially for large gridbox sizes, bilinear interpolation may not be physically persuasive, and the nearest neighbor approach is more compelling.

TRMM observations are temporally aggregated to the same periods as the station observations. As they do not cover the exact same periods, the first and last 3-hour TRMM observations are weighted by 0.5. For evaluation on different spatial scales, NWP forecasts and TRMM observations are aggregated to longitude–latitude boxes of $0.25° \times 0.25°$, $1° \times 1°$, and $5° \times 2°$. As propagation of precipitation systems is a potential error source and in an environment with predominantly westward movement of them, the largest box is tailored to assess NWP forecast quality without this potential source of error.

5.2.4 Consistency between TRMM and station observations

In light of the dry bias of TRMM observations, we evaluate the consistency of TRMM and station observations in our data sets. Specifically, we pair each station observation with the TRMM observation for the $0.25° \times 0.25°$ box that contains the station location. Figure 5.2 shows contingency tables of TRMM and station observations above and below 0.2 mm respectively, and two-dimensional frequency plots for TRMM and station observations above 0.2 mm, which is our threshold for the distinction between rain and no rain. We use this threshold irrespectively of the temporal and spatial aggregation at hand, with the results reported hereinafter being insensitive to this choice[1]. The Guinea Coast is moistest overall and has the highest fraction of warm rain events and isolated showers (Maranan et al., 2018; Young et al., 2018). In the absence of radar information, these rainfall events are often not detected by TRMM, while extensive cirrus is often misinterpreted as rainfall by TRMM (M. Maranan, personal communication, December 12, 2018). Both effects lead to the relatively large fractions of false positives and negatives in this region. West Sahel and East Sahel are drier than Guinea Coast, and reveal more correct negatives. While rainfall occurs less frequently in both Sahelian regions than in Guinea Coast, TRMM misinterprets also in these regions frequently occurring extensive cirrus as rainfall. For all regions the prevailing case is the one with both TRMM and the station reporting precipitation amounts below 0.2 mm. Among the disagreeing cases, the one with TRMM observing more than 0.2 mm and the station less than 0.2 mm is more frequent as suggested by the misinterpration of cirrus by TRMM and the mismatch in spatial scales between gridboxes and stations. The least squares regression lines in the two-dimensional frequency plots illustrate the dry bias of TRMM (e.g., Maggioni et al., 2016) relative to station observations when both report rain. Overall, the agreement between station and TRMM observations is fair. Disagreements of the magnitude and type seen here arise for reasons of differing coverage, spatial variability, and retrieval problems, among others, and are consistent with the extant literature (see, e.g., Roca et al., 2010; Engel et al., 2017; Maranan et al., 2018).

[1]Specifically, we checked thresholds from 0.0 mm to 1.0 mm, with minimal differences in findings.



	Stat ≤ 0.2	Stat > 0.2
TRMM > 0.2	9.1	17.6
TRMM ≤ 0.2	67.1	6.2

	Stat ≤ 0.2	Stat > 0.2
TRMM > 0.2	7.9	10.4
TRMM ≤ 0.2	77.3	4.4

	Stat ≤ 0.2	Stat > 0.2
TRMM > 0.2	16.0	28.5
TRMM ≤ 0.2	44.1	11.4

Figure 5.2: Comparison of 1-day accumulated station and TRMM observations of precipitation in monsoon seasons 2007–2014. The contingency tables contain the frequencies of TRMM and station observations below and above 0.2 mm respectively. The two-dimensional frequency plots show the joint distribution of TRMM and station observations above 0.2 mm, with the linear least squares line in red overlaid. Observations above 50 mm exist, but are very infrequent.

5.3 Results

Our annual evaluation period ranges from 1 May to 15 October, covering the wet period of the West African monsoon. The assessment of ECMWF ensemble forecasts is based on monsoon seasons 2007–2014, and for the other TIGGE subensembles we restrict the investigations according to availability as indicated in Table 4.1.

For verification against station observations, this yields more than 3,000, 6,000, and 12,000 forecast–observations pairs per monsoon season in East Sahel, West Sahel, and Guinea Coast. For verification against TRMM observations, we use 30 randomly chosen, non-overlapping boxes per region at $0.25° \times 0.25°$ and $1° \times 1°$ aggregation, and eight sites per region for $5° \times 2°$ longitude–latitude boxes. This covers substantial parts of the study region and results in about 5,000 forecast–observation pairs per monsoon season at the smaller aggregation levels, and well over 1,000 pairs at our highest level.

In Subsection 5.3.1, we study the skill of 1-day accumulated ECMWF raw and postprocessed ensemble precipitation forecasts in detail. Subsections 5.3.2 and 5.3.3 present results and highlight differences for longer accumulation times and spatially aggregated forecasts. Subsection 5.3.4 turns to results for all TIGGE

sub-ensembles, and we investigate the gain in predictability through inter-model variability using the RMM ensemble. In our (u)PIT histograms and reliability diagrams, we show results for the last available monsoon season only, given that operational systems continue to be improving (Hemri et al., 2014).

5.3.1 1-day accumulated ECMWF forecasts

Figure 5.3 shows (u)PIT histograms for 1-day accumulated raw and postprocessed ECMWF ensemble and EPC forecasts over West Sahel, East Sahel, and Guinea Coast. The histograms for the raw ensemble indicate strong underdispersion as well as a wet bias (panels a–c). At Guinea Coast, about 56% of the observations are smaller than the smallest ensemble member, a result that is robust across monsoon seasons. EMOS and BMA postprocessed forecasts generally are calibrated (panels g–l), as is EPC (panels d–f), except that the tails of the EMOS predictive distributions are too light as indicated by a too high rightmost bin. Statistical postprocessing also corrects for the systematically too high PoP values issued by the raw ECMWF ensemble. As shown in Figure 5.4, EMOS and BMA postprocessed PoP forecasts are reliable, but are hardly ever higher than 0.70. Generally, the postprocessed PoP forecasts have reliability and resolution similar to EPC.

Table 5.1 shows the mean BS, mean CRPS, and MAE for the various forecasts and regions, with the scores being averaged across monsoon seasons 2007–2014. We use a simple procedure to check whether differences in skill are stable across seasons. If a method has a higher (worse) mean score than EPC in all eight seasons, we mark the score $^{--}$; if it is judged worse in seven seasons, we put down a $^{-}$. Similarly, if a method has smaller (better) mean score than EPC in all seasons, we mark the score $^{++}$; if it performs better in seven seasons, we label $^{+}$ in the table. Viewed as a (one-sided) statistical test of the hypothesis of predictive skill equal to EPC, the associated tail probabilities or p-values are $1/2^8 = 0.0039\ldots$ and $(1 + 8)/2^8 = 0.035\ldots$ respectively. Clearly, the raw ECMWF ensemble underperforms relative to EPC, with $^{--}$ designations throughout. EMOS and BMA postprocessed forecasts perform at about the same level as EPC. For Guinea Coast, BMA receives $^{++}$ for CRPS and BS scores, but scores differences are small when compared to EPC. For the BS, the similar performance of postprocessed and EPC forecasts stems from the fact that not only do postprocessed and EPC forecasts show similar reliability but also similar resolution, as seen from the inset histograms in panels d–l of Figure 5.4.

The Murphy diagrams in the top row of Figure 5.5 corroborate these findings. For 1-day precipitation occurrence, decision makers will mostly prefer the climatological reference EPC over the raw ECMWF ensemble, and only some decision makers will have a slight preference for EMOS or BMA postprocessed forecasts, as compared to EPC. Further light on theses issues is shed by the ROC diagrams in the bottom row of the figure. EMOS and BMA PoP forecasts can be interpreted as recalibrated raw ensemble probabilities, and so it is not surprising that for West Sahel and East Sahel, raw and postprocessed forecasts show essen-

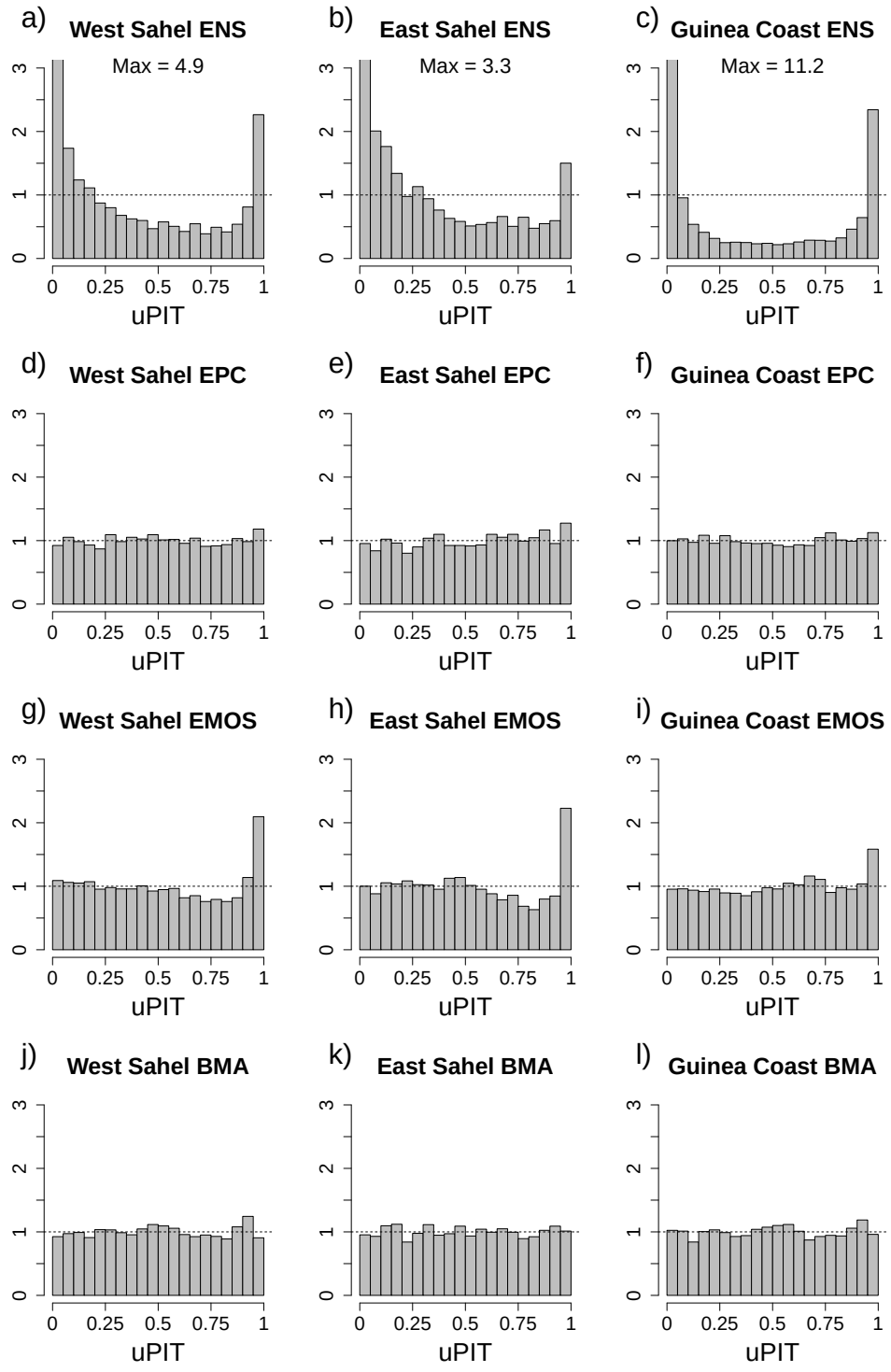


Figure 5.3: Unified PIT (uPIT) histograms for raw ECMWF ensemble, EPC, and EMOS and BMA postprocessed forecasts of 1-day accumulated precipitation in monsoon season 2014, verified against station observations. Histograms are cut at a height of 3, with the respective maximal height noted. The dashed line indicates the uniform distribution that corresponds to a calibrated forecast. © Copyright 2018 AMS.

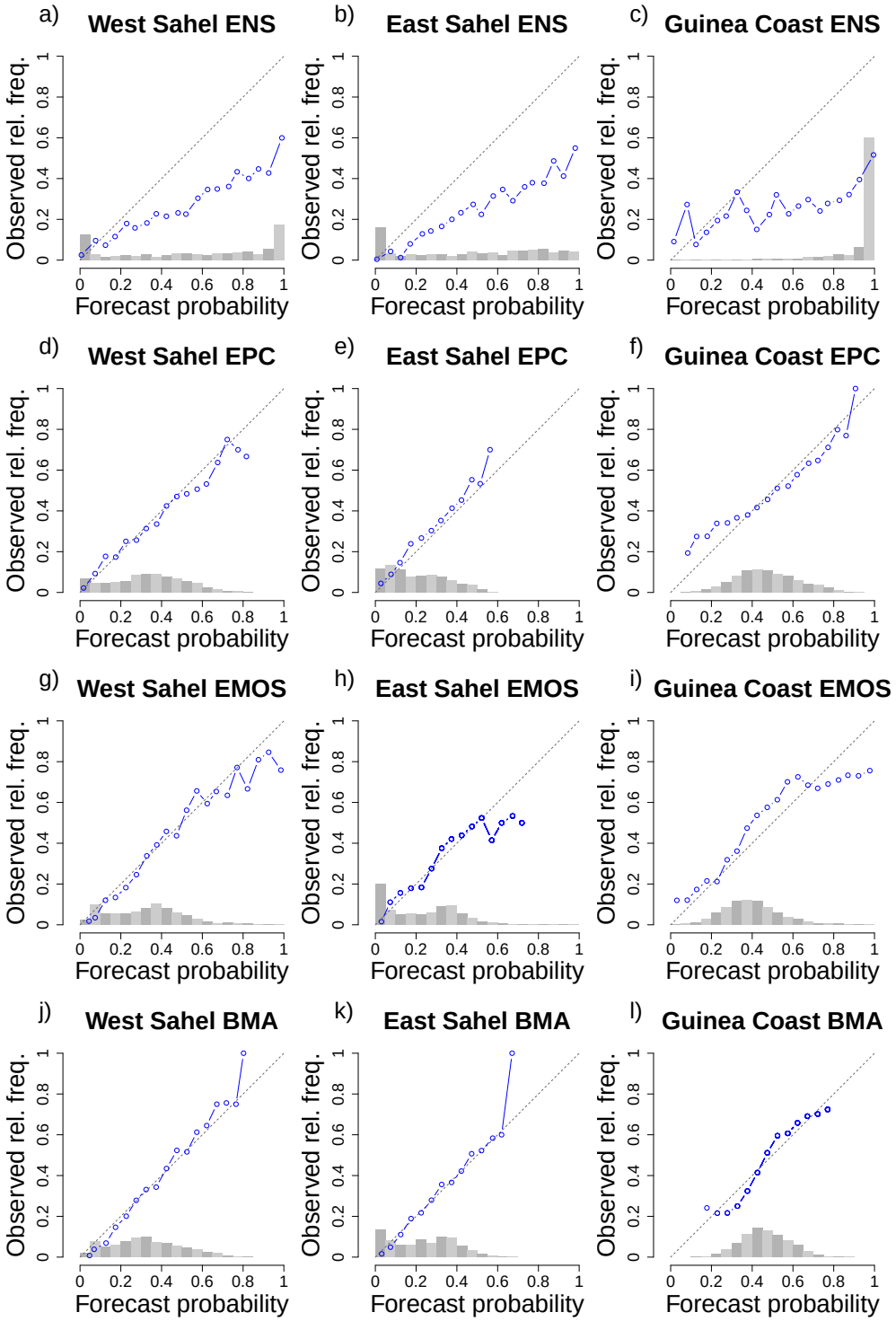


Figure 5.4: Reliability diagrams for raw ECMWF ensemble, EPC, and EMOS and BMA postprocessed forecasts of 1-day accumulated precipitation in monsoon season 2014, verified against station observations. The diagonal indicates perfect reliability, and the histograms show the relative frequencies of the PoP forecast values. © Copyright 2018 AMS.

Table 5.1: Mean BS at a threshold of 0.2 mm, mean CRPS, and MAE for raw ECMWF ensemble, EPC, and EMOS and BMA postprocessed forecasts of 1-day accumulated precipitation in monsoon seasons 2007–2014, verified against station observations. If a method has a higher (worse) respectively lower (better) mean score than EPC in all eight seasons, the score is marked $^{--}$ respectively $^{++}$; if it performs worse respectively better than EPC in seven seasons, the score is marked $^{-}$ respectively $^{+}$.

	BS			CRPS			MAE		
	West Sahel	East Sahel	Guinea Coast	West Sahel	East Sahel	Guinea Coast	West Sahel	East Sahel	Guinea Coast
ENS	$^{--}$0.32	$^{--}$0.32	$^{--}$0.48	$^{--}$4.50	$^{--}$2.63	$^{--}$6.99	$^{--}$5.36	$^{--}$3.13	$^{--}$8.39
EPC	0.19	0.15	0.23	3.75	2.08	5.28	4.60	2.38	6.57
EMOS	0.19	0.15	0.23	3.75	2.15	$^{+}$5.25	4.65	$^{--}$2.45	6.60
BMA	$^{+}$0.18	0.15	$^{++}$0.22	3.71	2.07	$^{++}$5.20	4.58	2.38	6.53

tially the same discrimination skill, at a level that is slightly superior to EPC. For Guinea Coast, EMOS and BMA have considerably higher AUC than the raw ensemble, due to the extreme concentration of the raw ensemble probabilities at very high levels, as illustrated in panel c of Figure 5.4. In contrast, the Murphy curves are sensitive to calibration and show marked differences between raw and postprocessed forecasts. Overall, these are sobering results, as they suggest that over northern tropical Africa ECMWF 1-day accumulated precipitation forecasts are hardly of practical use.

What could be possible reasons for the poor performance of the raw forecasts? A number of recent studies have shown that the use of convective parametrization is a first-order error source for realistically representing precipitation, cloudiness, wind and even the regional-scale monsoon circulation in West Africa together with their respective diurnal cycles (e.g., Pearson et al., 2014; Marsham et al., 2013; Birch et al., 2014; Pantillon et al., 2015). Based on these results, and given that all models we investigate use convective schemes, we suspect this aspect to be a major cause of the poor performance we find. A visual comparison of 1-day accumulated precipitation forecasts from ECMWF HRES and TRMM shows that rainfall structures in the model tend to be too widespread and too light lacking signs of mesoscale organization (see Figure 5.10 for an example).

Inspection of raw ensemble data suggests that for both station and TRMM observations, agreement between forecasts and observations is modest at best. Many observed precipitation events are either not predicted at all, are strongly underpredicted, or are predicted by (almost) all ensembles members (with varying amounts of precipitation), yet are not observed (see Figure 5.11 for an illustrative example). The last point may indicate a form of triggering that the model responds to too easily.

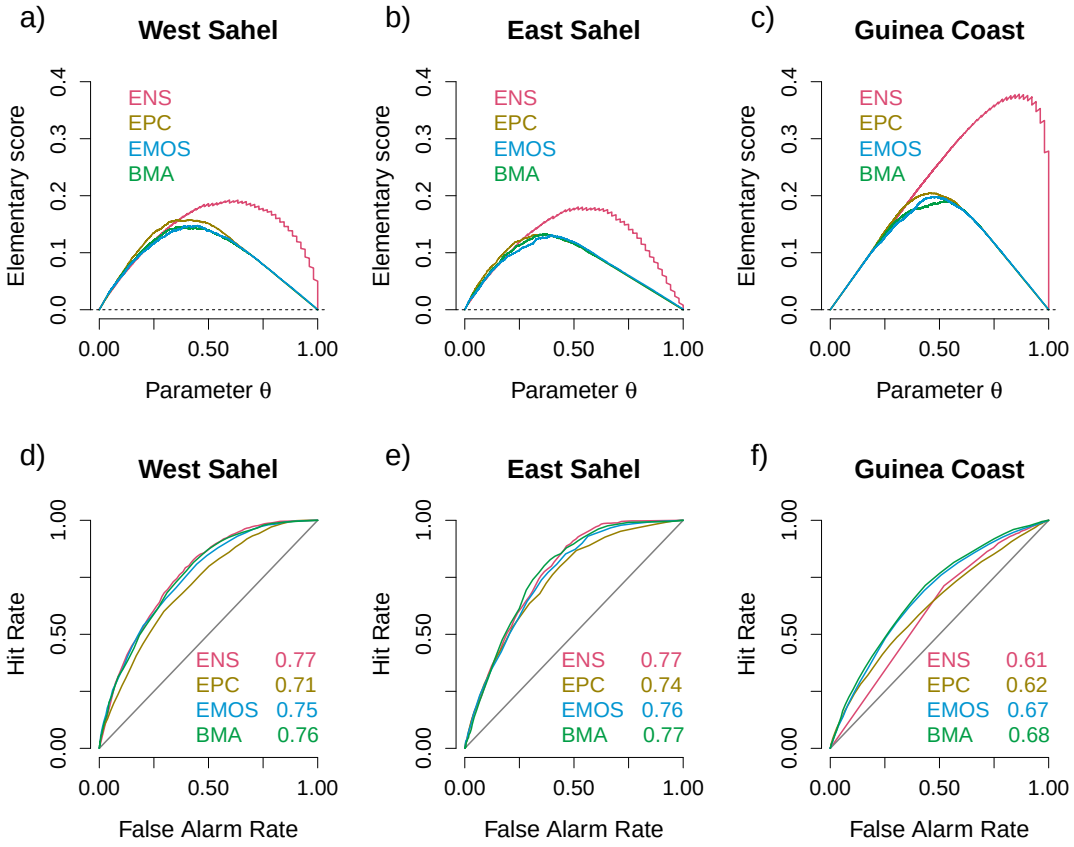


Figure 5.5: Murphy diagrams and ROC curves (with respective AUC values) for raw ECMWF ensemble (ENS), EPC, and EMOS and BMA post-processed 1-day accumulated PoP forecasts in monsoon season 2014, verified against station observations. © Copyright 2018 AMS.

5.3.2 Longer accumulation times

One might expect NWP precipitation forecasts to improve relative to EPC at longer accumulation times, as the main focus in forecasting shifts from determining time and location of initiation and subsequent propagation of convection towards determining regions with enhanced or reduced activity, based on large-scale conditions. Longer lead times might also lead to growth in differences between perturbed members, and thus reduce raw ensemble underdispersion.

However, the PIT histogram in Figure 5.6a indicates only slight, if any, improvement in calibration for raw ECMWF 5-day accumulated precipitation forecasts over West Sahel, and the results for the other regions are similar (not shown). Raw ensemble reliability improves at longer accumulation times, verified against either station observations in panel b), or $5° \times 2°$ TRMM observations in panels c) and d), though at a loss of resolution.

Table 5.2 uses the same approach as Table 5.1, but the scores are now for 5-day accumulated precipitation. The raw ECMWF ensemble still underperforms

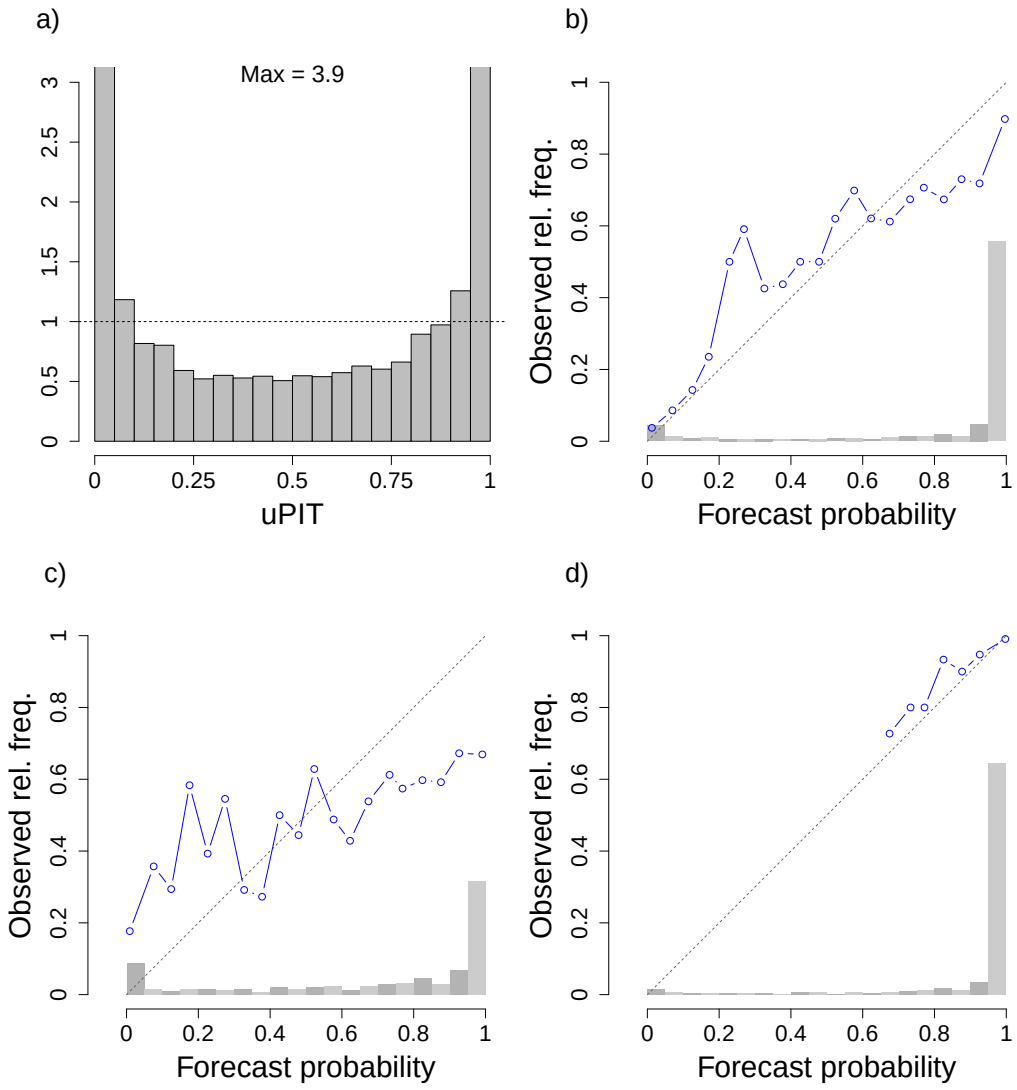


Figure 5.6: Calibration and reliability of raw ECMWF ensemble forecasts over West Sahel in monsoon season 2014 at 1- and 5-day accumulations. a) (u)PIT histogram and b) reliability diagram for 5-day accumulated precipitation, verified against station observations. Panels c) and d) show reliability diagrams for 1- and 5-day accumulated precipitation, verified again $5° \times 2°$ aggregated TRMM observations. Same approach as in Figures 5.3 and 5.4.

relative to EPC. EMOS and BMA postprocessed forecasts outperform EPC only slightly, with the differences in scores being small and generally not stable across monsoon seasons as indicated by only few $^{+}$ or $^{++}$. Despite the change in the underlying forecast problem, even postprocessed ECMWF ensemble forecasts are generally not superior to EPC.

Table 5.2: Mean BS, mean CRPS, and MAE for raw ECMWF ensemble, EPC, and EMOS and BMA postprocessed forecasts of 5-day accumulated precipitation in monsoon seasons 2007–2014, verified against station observations. Same approach as in Table 5.1.

	BS			CRPS			MAE		
	West Sahel	East Sahel	Guinea Coast	West Sahel	East Sahel	Guinea Coast	West Sahel	East Sahel	Guinea Coast
ENS	0.14	$^{--}$0.25	$^{--}$0.10	$^{--}$12.80	$^{--}$8.42	$^{--}$19.69	16.23	$^{--}$10.76	$^{-}$24.41
EPC	0.12	0.16	0.08	11.63	7.07	16.54	16.15	9.56	22.98
EMOS	0.13	0.16	$^{-}$0.08	11.62	7.34	16.44	$^{+}$15.99	9.96	22.74
BMA	$^{+}$0.11	$^{++}$0.15	0.08	$^{++}$11.47	$^{+}$6.94	16.33	$^{+}$16.07	$^{+}$9.45	22.92

5.3.3 Spatially aggregated observations

For the assessment of forecast skill at larger spatial scales, we focus on ECMWF raw and BMA postprocessed ensemble forecasts over West Sahel, evaluated by BS and CRPS. This is due to the similarities in CRPS and MAE results, better performance of BMA compared to EMOS in many instances, and results for West Sahel that are as good for BMA postprocessed forecasts as for East Sahel, and better than for Guinea Coast.

The use of spatially aggregated TRMM observations avoids problems of point to pixel comparisons, and at higher aggregation we can assess forecast quality with minimal error due to the propagation of convective systems. The dry bias of TRMM disadvantages the raw ensemble compared to EPC and postprocessed forecasts, but does not hinder assessments regarding systematic forecast errors. As illustrated in Figure 5.6c, 1-day PoP forecasts from the raw ECMWF ensemble remain unreliable even at the $5^\circ \times 2^\circ$ gridbox scale. It is only under large scales and longer accumulation times simultaneously, when precipitation occurs almost invariably, that raw ensemble PoP forecasts become reliable (panel d).

Table 5.3 shows mean BS and CRPS scores at various spatial aggregations for 1-day precipitation accumulation, verified against TRMM observations. The raw ECMWF ensemble forecast is inferior to EPC at all resolutions, and in every single region and season. BMA postprocessed forecasts outperform EPC across aggregation scales, and in every single region and season, but the improvement relative to EPC remains small.

Table 5.3: Performance of spatially aggregated raw ECMWF ensemble, EPC, and BMA postprocessed forecasts of 1-day accumulated precipitation in monsoon seasons 2007–2014, verified against TRMM gridbox observations. Same approach as in Table 5.1.

	TRMM 0.25° × 0.25° / 1 d						TRMM 1° × 1° / 1 d			TRMM 5° × 2° / 1 d		
	BS			CRPS			CRPS			CRPS		
	West Sahel	East Sahel	Guinea Coast	West Sahel	East Sahel	Guinea Coast	West Sahel	East Sahel	Guinea Coast	West Sahel	East Sahel	Guinea Coast
ENS	$^{--}$0.30	$^{--}$0.23	$^{--}$0.48	$^{--}$2.29	$^{--}$1.44	$^{--}$4.03	$^{--}$2.24	$^{--}$1.56	$^{--}$4.43	$^{--}$1.95	$^{--}$1.53	$^{--}$4.22
EPC	0.19	0.14	0.23	1.07	0.57	1.35	0.94	0.58	1.36	0.81	0.49	1.07
BMA	$^{++}$0.17	$^{++}$0.13	$^{++}$0.21	$^{++}$1.03	$^{++}$0.55	$^{++}$1.29	$^{++}$0.89	$^{++}$0.55	$^{++}$1.28	$^{++}$0.76	$^{++}$0.45	$^{++}$0.95

5.3.4 TIGGE sub-ensembles and RMM ensemble

In addition to the ECMWF EPS, which we have studied thus far, the TIGGE database contains several more operational sub-ensembles, as listed in Table 4.1. Figure 5.7 shows PIT histograms for the various sub-ensembles and the RMM ensemble for 1-day accumulated precipitation forecasts over West Sahel. All TIGGE sub-ensembles exhibit underdispersion and wet biases, though in strongly varying degrees. The MF raw ensemble is the most underdispersive ensemble and more than every second observation is smaller than the smallest MF ensemble member. The RMM ensemble is better calibrated than most of its contributors, while the Meteorological Service of Canada (MSC) model is the most calibrated sub-ensemble of TIGGE.

Figure 5.8 displays BS and CRPS skill relative to EPC for raw and BMA postprocessed TIGGE sub-ensemble and RMM ensemble forecasts in 2007–2014, verified against station observations. All raw ensembles underperform relative to EPC, in part drastically so. For most sub-ensembles, a temporal improvement in skill is visible, with monsoon seasons 2011–2014 revealing higher skill than 2007–2010. Of all TIGGE sub-ensembles, MSC typically is the most skillful one, while MF and CPTEC perform quite poor. But also renowned models such as NCEP, ECMWF, or UKMO struggle to issue skillful forecasts. Across 2007–2014, the ECMWF model reveals the highest improvement in both BS and CRPS skill. Postprocessing by BMA increases forecast quality. The ECMWF, KMA, NCEP, and UKMO ensembles yield the best postprocessed forecasts, exhibiting small positive skill relative to EPC for most monsoon periods. The BMA postprocessed RMM ensemble outperforms all sub-ensembles as well as EPC, but the improvement is small. Figure 5.9 displays the BMA weights that the 15 contributors of the RMM ensemble attain (see (4.4)). The mean perturbed forecasts from the ECMWF, UKMO, and NCEP ensembles are the top three contributors to the BMA postprocessed RMM forecast, while the best control run only receives a weight of about 0.05.

In further experiments, we have studied raw and postprocessed TIGGE sub-ensemble and RMM ensemble forecasts at accumulation times up to 5 days and spatial aggregations up to 5° × 2° gridboxes in TRMM as already presented for

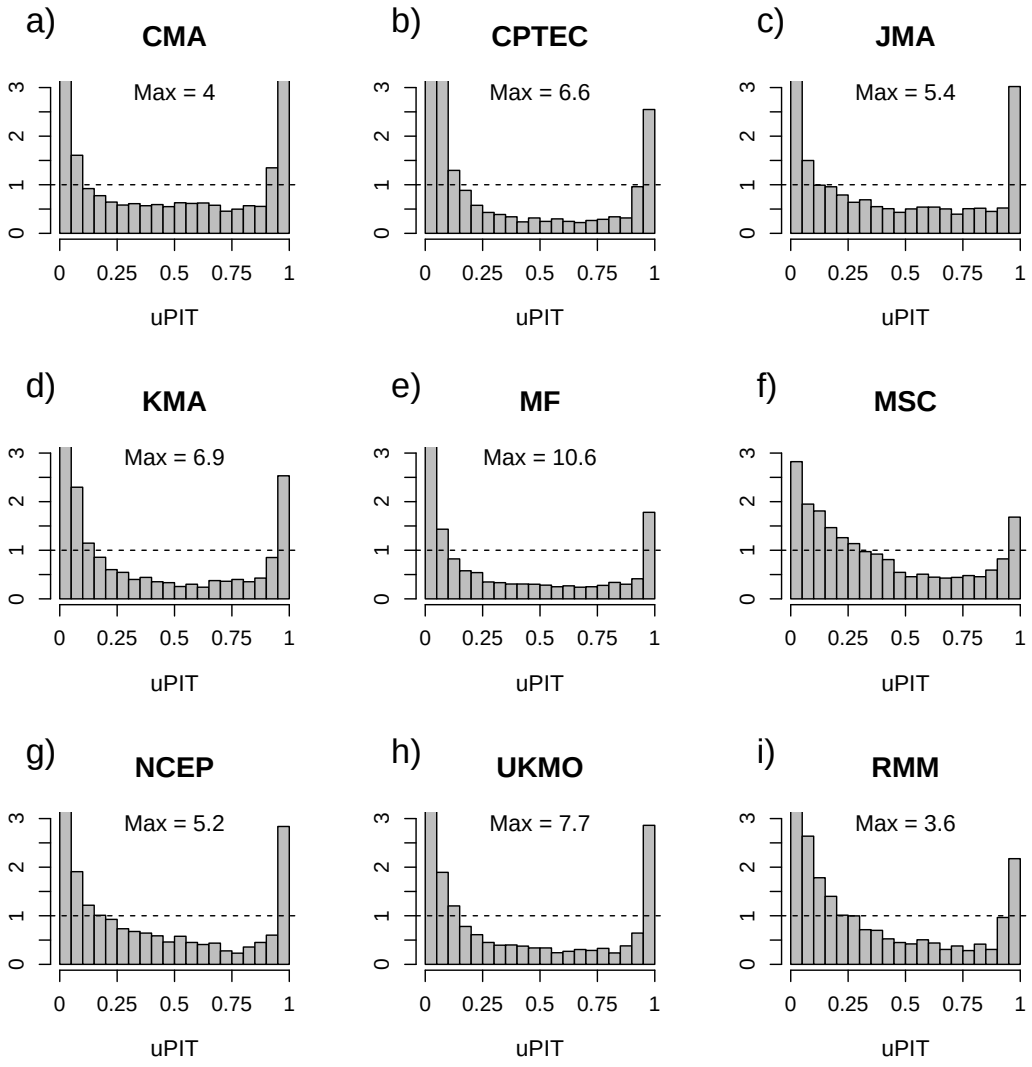


Figure 5.7: (u)PIT histograms for raw TIGGE sub-ensemble and raw RMM ensemble forecasts of 1-day accumulated precipitation over West Sahel in monsoon season 2013, verified against station observations. Same approach as in Figure 5.3. © Copyright 2018 AMS.

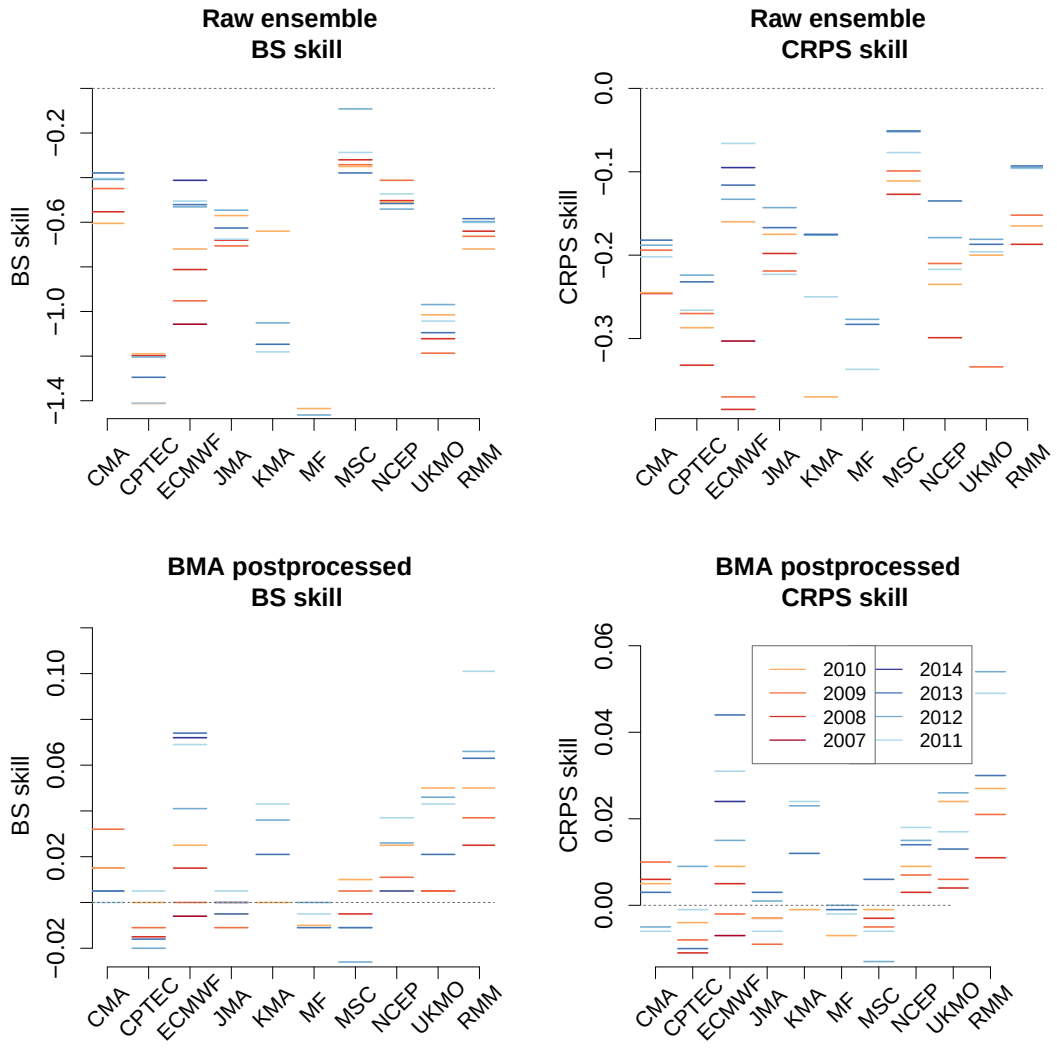


Figure 5.8: BS and CRPS skill for raw and BMA postprocessed TIGGE sub-ensemble forecasts of 1-day accumulated precipitation over West Sahel in monsoon seasons 2007–2014, verified against station observations. Skill equal to EPC is indicated by the dashed line. © Copyright 2018 AMS.

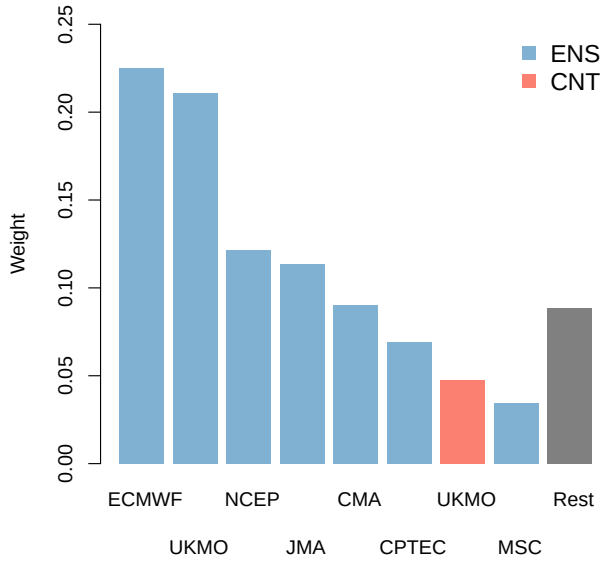


Figure 5.9: BMA weights of RMM components for 1-day accumulated precipitation forecasts over West Sahel trained against station observations, averaged over monsoon seasons 2008–2013. Mean perturbed forecasts (ENS) and control runs (CNT) are distinguished by the color of the respective bar.

ECMWF in Subsection 5.3.3. Our findings generally remain unchanged. The raw ensemble forecasts never reach the quality of the climatological reference EPC. After postprocessing with BMA, the ECMWF ensemble typically becomes the best performing TIGGE sub-ensemble, showing slightly better scores than EPC when verified against TRMM observations, at all spatial aggregations. The BMA postprocessed RMM forecast depends heavily on the ECMWF mean perturbed forecast, and is superior to both EPC and the BMA postprocessed sub-ensemble.

5.4 Discussion

In a first-ever thorough verification study over northern tropical Africa, the quality of operational ensemble precipitation forecasts from different NWP centers was assessed for several years, accumulation periods, and for station and spatially aggregated satellite observations. All raw ensembles exhibit calibration problems in form of underdispersion and biases, and are unreliable at high PoP forecast values. They have lower skill than the climatological reference EPC for the prediction of occurrence and amount of precipitation, with the underperformance being stable across monsoon seasons.

After correcting for systematic errors in the raw ensemble through statistical postprocessing, the ensemble forecasts become reliable and calibrated, but only

few are slightly superior to EPC. While further developments of both EMOS and BMA might be feasible (see, e.g., Fortin et al., 2006; Scheuerer and Hamill, 2015), and training sets could be augmented by using reforecast data (e.g., Di Giuseppe et al., 2013), the respective benefits are likely to be incremental at this time, although as the raw ensemble performance improves over time, they might become considerable. Not surprisingly, forecast skill tends to be highest for long accumulation times and large spatial aggregations. Overall, raw ensemble forecasts are of no use for the prediction of precipitation over northern tropical Africa, and even EMOS and BMA postprocessed forecasts have little added value compared to EPC.

What are the reasons for this rather disappointing performance of state-of-the-art global EPSs? For 1-day accumulated precipitation forecasts, the ability of an NWP model to resolve the details of convective organization is essential. As all global EPSs use parameterized convection, this likely limits forecast skill. The fact that even postprocessed 1-day accumulated ensemble forecasts exhibit no skill relative to EPC, implies that ensembles cannot translate information about the current atmospheric state (e.g., tropical waves or influences from the extratropics) into meaningful impacts regarding the occurrence or amount of precipitation. This is robust for verification against station as well as spatially aggregated satellite observations, and can therefore not be explained by propagation errors.

For longer accumulation times and larger spatial aggregations, the large-scale circulation has a much stronger impact on convective activity, which should weaken the limitation through convective parameterization. The skill of 5-day accumulated precipitation forecasts, however, increases only slightly, if at all, compared to 1-day accumulated forecasts. The most likely reason for this is that squall lines have feedbacks on the large-scale circulation, which are not realistically represented in global NWP models either. Marsham et al. (2013) find that the large-scale monsoon state in (more realistic) simulations with explicit convection differs quite pronouncedly from runs with parameterized convection, even when using the same resolution of 12 km. In the explicit-convection simulation, greater latent and radiative heating in the Sahel weakens the monsoon flow, delays the diurnal cycle, and convective cold pools provide an essential component to the monsoon flux. We suspect that some or all of these effects are misrepresented in global EPS forecasts.

The fact that EPS precipitation forecasts are so poor over northern tropical Africa is a strong demonstration of the complexity of the underlying forecast problem. An interesting question in this context is whether poor predictability in the tropics is exclusively confined to northern Africa, where AEWs provide favorable conditions for convective organization into MCSs ahead of the trough.

Furthermore, the lack of skill motivates complementary approaches to predicting precipitation over this region. Little et al. (2009) compare operational NCEP ensemble, climatological, and statistical forecasts for stations in the Thames Valley, United Kingdom. They note that NCEP forecasts outperform climatological forecasts, but demonstrate that statistical forecasts, solely based on past observations, can outperform NCEP forecasts by exploiting spatio-temporal depen-

dencies. These also exist over northern tropical Africa and some additional predictability may stem from large-scale drivers such as convectively-coupled waves. Fink and Reiner (2003) note a coupling of the initiation of squall lines to AEWs and Wheeler and Kiladis (1999) the influence of large-scale tropical waves, such as Kelvin and equatorial Rossby waves or the Madden-Julian oscillation, on convective activity. Pohl et al. (2009) confirm the relation between the Madden-Julian oscillation and rainfall over West Africa and Vizy and Cook (2014) demonstrate an impact of extratropical wave trains on Sahelian rainfall. Statistical models based on spatio-temporal characteristics of rainfall and extended by such large-scale predictors seem a promising approach to improve precipitation forecasts over our study region, and we expect such forecasts to outperform climatology. This approach will be explored in future work, and results of a pilot study for northern tropical Africa are presented in Chapter 7.

As discussed in Section 5.3.1, we suspect convective parametrization to be a major cause of the low quality of model-based forecasts here. Therefore it would be interesting to test ensembles of convection-permitting NWP model runs, ideally in combination with ensemble data assimilation, but the computational costs are high, and it will take time until a multi-year database will become available for validation studies. Alternatively, it could be tested whether systematic improvements to convection schemes (e.g., Bechtold et al., 2014) do in fact positively impact on ensemble forecast quality. Given the growing socioeconomic impact of rainfall in northern tropical Africa with its rain-fed agriculture, statistical and statistical-dynamical approaches should be developed in parallel in order to improve the predictability of rainfall in this region.

Appendix 5.A Quality control for rainfall observations within KASS-D

Rainfall exhibits extremely high spatial and temporal variability, which hinders automated quality checks applicable to other meteorological variables such as temperature or pressure. For precipitation, Fiebrich and Crawford (2001) suggest a range and a step test only. The global range of station observed 1-day accumulated precipitation is from 0 mm to 1,825 mm. All KASS-D observations passed this test. The step test checks if the difference of neighboring 5-minute accumulated precipitation is smaller than 25 mm. For 1-day accumulated precipitation tests of this type are not meaningful, nor are the persistence tests used by Pinson and Hagedorn (2012) for wind speed. However, the site-specific climatological distributions of precipitation accumulation should be right-skewed, i.e., the median should be smaller than the mean, and in the tropics they should have a point mass at zero (Rodwell et al., 2010). As noted, we only consider stations with more than 80% available observations in any of the monsoon seasons, and all 132 stations thus selected passed these tests.

Appendix 5.B Consistency of ECMWF forecasts and verifying observations

Figure 5.10 displays maps of 1-day accumulated precipitation as forecasted by the ECMWF HRES run and observed by TRMM on 14 July 2014. Precipitation over Guinea and Mali is well predicted in terms of location, even though the organization seems less well captured. However, most of the forecasted precipitation over Nigeria did not to materialize, and in the East Sahel region precipitation occurred over Sudan rather than the Ethiopian Highlands.

Figure 5.11 displays time series of 1-day accumulated ECMWF precipitation forecasts along with the respective station and TRMM observations. The titles of the panels indicate the WMO station number or the longitude and latitude coordinates of the center of the considered $0.25° \times 0.25°$ TRMM pixel. For both types of observations, there is a modest degree of agreement between forecasts and observations. However, many precipitation events are either not predicted at all, are strongly underpredicted, or are predicted by (almost) all ensembles members (with varying amounts of precipitation), yet do not occur.

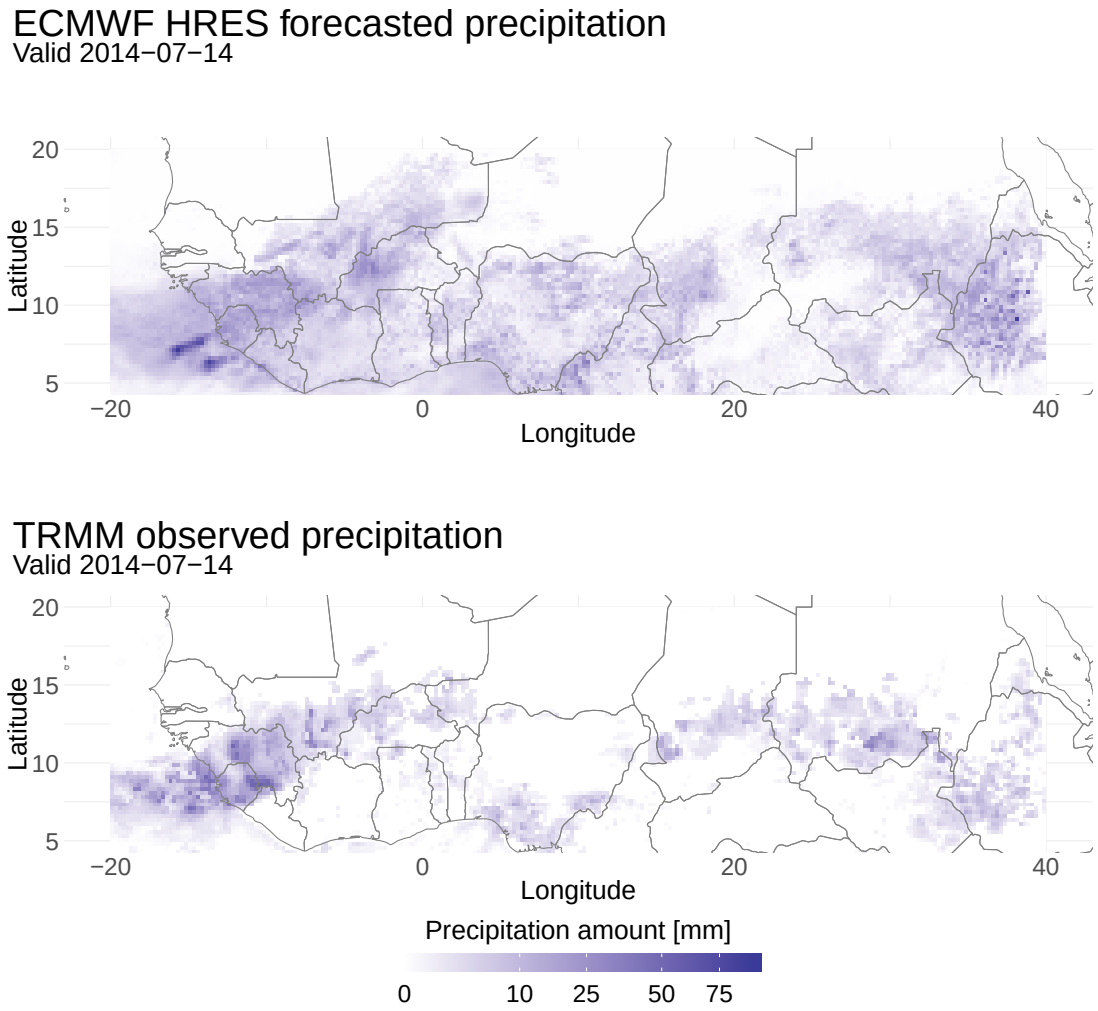


Figure 5.10: 1-day accumulated precipitation on 14 July 2014 as forecasted by the ECMWF HRES run and observed by TRMM at a resolution of $0.25° \times 0.25°$. © Copyright 2018 AMS.

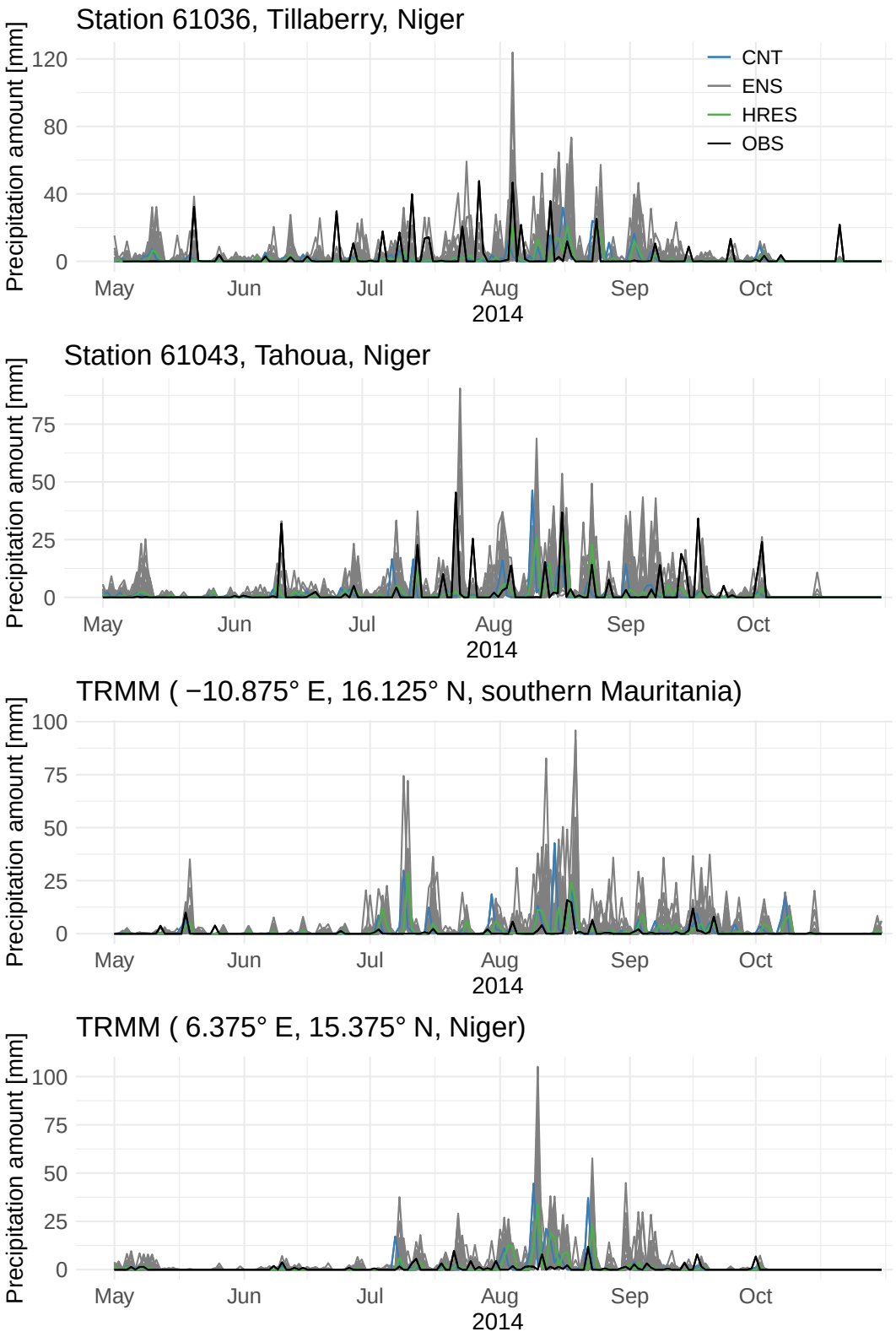


Figure 5.11: ECMWF ensemble forecasts over West Sahel in 2014 along with respective station or $0.25° \times 0.25°$ TRMM observations. The HRES and CNT runs are plotted along with the 50 perturbed members.

6 Skill of global raw and postprocessed ensemble predictions of rainfall in the tropics

In this chapter, we systematically evaluate for the first time NWP ensemble forecast skill for amount and occurrence of precipitation as well as the occurrence of extreme precipitation events in ten climatic regions in the tropics. We evaluate ensemble forecasts from ECMWF and MSC for accumulation periods of 1–5 days against TRMM observations across the period 2009–2017. In order to evaluate the full potential of ensemble forecasts, we apply the state-of-the-art statistical postprocessing technique EMOS and compare raw and postprocessed forecast against EPC.

6.1 Introduction

Throughout the tropics, forecasts of precipitation have a multitude of users, both at the short range and at seasonal timescales. Despite this, little is known about the quality of current NWP ensemble forecasts for precipitation at accumulation periods of one to a few days. Haiden et al. (2012) introduced the stable equitable error in probability space (SEEPS) score that classifies precipitation forecasts and observations into three categories based on the local climate. It allows to assess deterministic forecast quality for different climatic regions and Haiden et al. (2012) note that "SEEPS scores at forecast day 1 in the tropics are similar to those at day 6 in the extratropics". In Chapter 5, we found little to no skill in 1–5 day accumulated precipitation forecasts from ten global NWP EPSs for northern tropical Africa. These results are robust under temporal and spatial aggregation and point to fundamental problems in predicting precipitation in this region, and potentially in the tropics as a whole. In contrast, Webster (2013) reports relatively good forecasts of precipitation for southern Asia up to ten days ahead.

Without an assessment of the quality of accumulated precipitation forecasts from current NWP ensembles, the further improvement of NWP models is hindered. The ultimate aim of this chapter is to provide a detailed analysis of our current ability to predict rainfall, rainfall occurrence, and extreme rainfall for the tropics and at a regional level by assessing global raw and postprocessed forecasts from two major NWP centers. We examine accumulation periods of 1–5 days for

the period 2009–2017 and verify against satellite-based gridded precipitation observations for 21,600 gridboxes between 30°S and 30°N that allow a fine-grained assessment.

Section 6.2 introduces the analyzed ensemble forecasts and satellite observations as well as the climatic regions. Section 6.3 presents the results of our investigations and relies on the ECMWF EPS as key exemplar. We first assess calibration and reliability of ECMWF raw and postprocessed forecasts before considering their skill for the prediction of amount and occurrence of precipitation and the occurrence of extreme precipitation. Subsequently, we compare these results to ensemble forecasts from MSC and analyze improvement over the period 2009–2017. Section 6.4 discusses potential reasons for and implications of our results.

6.2 Data

6.2.1 Forecasts

Due to its high quality, the ECMWF EPS serves as key exemplar for the analysis of NWP ensemble forecast skill for accumulated precipitation. Its setup and properties are discussed in Chapter 4. For an additional evaluation, we rely on the MSC EPS. It is among the best EPSs for predictions of accumulated precipitation in northern tropical Africa (see Figure 5.8) and one of the leading EPSs worldwide. The forecast quality of both EPSs can be monitored in quasi-real time at the WMO Lead Centre on Verification of Ensemble Prediction Systems website `http://epsv.kishou.go.jp/EPSv`.[1] It displays average scores for standard atmospheric variables for the tropical belt between 20°S and 20°N, and northern and southern hemisphere extratropics. MSC ensemble forecasts are accessible via the TIGGE archive with a spatial resolution of $0.5° \times 0.5°$. For both models, we rely on forecasts initialized at 00 UTC.

6.2.2 Observations

For a spatially consistent and complete verification of NWP ensemble forecast quality, we rely on the TRMM 3B42 gridded data set (see Section 5.2 for further details). It is regarded the best available satellite precipitation product (see, e.g., Maggioni et al., 2016) despite its mediocre performance in detecting rainfall in complex terrain and semiarid areas, and its dry bias that is particularly large for light rain events (Huffman et al., 2007). If nearby gauge observations are available, they are used to calibrate TRMM estimates based on monthly accumulation sums. The result is an observational data set with full spatial coverage and only a small bias on monthly scales. Over oceans, calibration is not possible and satellite-based measurements used in the TRMM algorithm are less capable in detecting precipitation. This holds in particular for regions where most rain

[1] On this webpage, the MSC is denoted as Canadian Meteorological Centre (CMC).

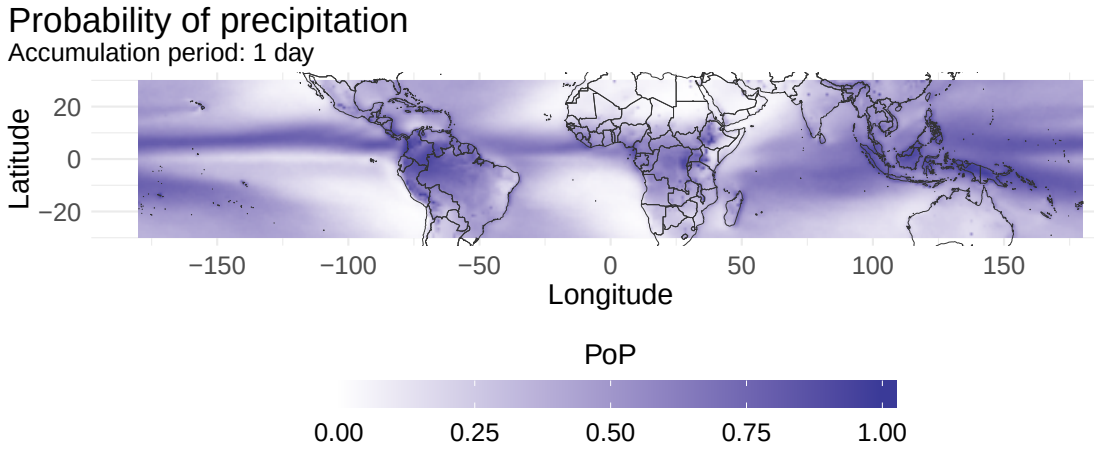


Figure 6.1: Climatological PoP. A threshold of 0.2 mm is used to determine the occurrence of precipitation for an accumulation period of 1 day.

events are light (Huffman et al., 2007). Figure 6.1 displays the climatological PoP for an accumulation period of one day, and several oceanic and continental deserts with very low climatological PoP are visible. In particular for the oceanic deserts west of South America and southern Africa, (rarely occurring) rainfall is almost exclusively light as suggested by panel a) of Figure 11 in Nesbitt et al. (2006) and TRMM observations for these regions are least reliable.

6.2.3 Data preprocessing

We apply the exact same procedures for data preprocessing as in Chapter 5, and aggregate TRMM observations and ensemble forecasts to a resolution of $1^\circ \times 1^\circ$ and accumulation periods of one to five days. This yields a total of 21.600 gridboxes for the tropics between 30°S and 30°N and allows for a fine-grained investigation of NWP EPS precipitation forecast quality.

6.2.4 Köppen-Geiger climates

We divide the tropics into Köppen-Geiger climates by continents for an assessment of forecast quality at a regional level. The Köppen-Geiger climate classification (Köppen, 1900; Geiger, 1961) uses five main climates, and subgroups within each climate that are defined by seasonal precipitation patterns. Kottek et al. (2006) provide an updated Köppen-Geiger climate classification with a resolution of $0.25^\circ \times 0.25^\circ$, available at `http://koeppen-geiger.vu-wien.ac.at/present.htm`. We rely on the main climates only and merge Continental (D) and Polar (E) climates as there only exist 6 and 91 gridboxes with continental and polar climates at a resolution of $1^\circ \times 1^\circ$ in the tropics. In the following, we call the resulting areas "Alpine climates". For the most frequent climates Tropical (A) and Arid (B), we use an additional stratification by conti-

nents such that, e.g., Africa is divided into northern arid Africa, tropical Africa, and southern arid Africa. This yields a total of ten climatic regions, displayed by color in Figure 6.2.

6.3 Results

6.3.1 Calibration and reliability of the ECMWF ensemble

Figure 6.2 displays in the top panel PIT histograms for 1-day accumulated precipitation forecasts by the ECMWF ensemble. Here and in the following, the distribution into climatic regions follows the Köppen-Geiger climates and their stratification by continents introduced in Subsection 6.2.4. The PIT histograms reveal that for all regions ECMWF raw ensemble forecasts are strongly underdispersive, and in many regions more than 40% of all observations are smaller than the smallest ensemble member as indicated by a leftmost bar having a height of more than 8. The bottom panel of Figure 6.2 displays the spatial distribution of the (scaled) discrepancy measure between the ECMWF raw ensemble and perfectly calibrated forecasts as defined by Berrocal et al. (2007). It attains values between zero and one, where lower values indicate better calibration. Typically, the ECMWF ensemble is better calibrated over land than over ocean, but it is not well calibrated anywhere in the tropics. Regions of particularly low calibration are the oceanic deserts, e.g., west of South America and southern Africa. The lack of calibration in ECMWF ensemble forecasts is robust across accumulation times from 1–5 days (not shown).

The top panel of Figure 6.3 displays reliability diagrams for 1-day accumulated PoP forecasts by the ECMWF raw ensemble. Raw ensemble forecasts for the occurrence of precipitation are generally overconfident and unreliable as expressed by too frequent PoP forecasts of very high probabilities for rainfall occurrence and non-occurrence and less frequent realizations of the predicted event. Especially over complex terrain as found in alpine climates, the ECMWF EPS struggles to produce reliable forecasts. As threshold for the occurrence of precipitation, we rely here and in the following on 0.2 mm irrespective of the accumulation period, but note that our results change only minimally under different choices of the threshold such as 1 mm.

After statistical postprocessing, ECMWF forecasts are fairly well calibrated as is EPC, even though small deviations from a uniform distribution typically remain (not shown). The bottom panel of Figure 6.3 displays reliability diagrams for postprocessed ECMWF PoP forecasts for an accumulation period of 1 day. As statistical postprocessing corrects for systematic overconfidence and unreliability, EMOS postprocessed forecasts are more reliable than raw ensemble forecasts, even though they are slightly underconfident. The increase in reliability is partially achieved at the cost of a lower resolution of postprocessed forecasts as displayed by the inset histograms in Figure 6.3. The EPC forecast is also reliable with a resolution comparable to that of ECMWF postprocessed forecasts.

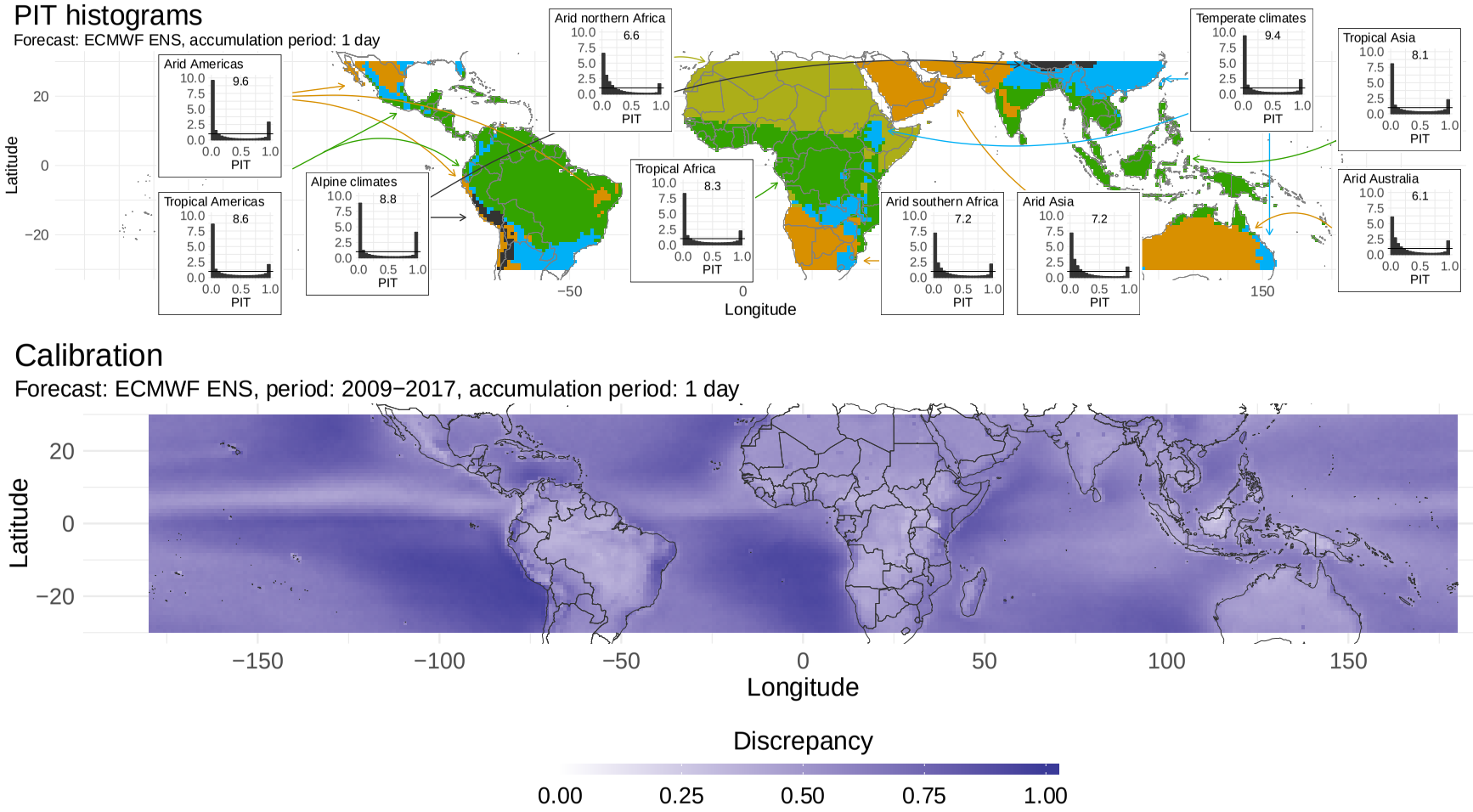


Figure 6.2: Calibration of ECMWF raw ensemble forecasts for 1-day accumulated precipitation. The top panel displays PIT histograms with 20 bins for the ten Köppen-Geiger climates. The height of the highest bin is noted. The bottom panel shows the spatial distribution of the discrepancy between ECMWF raw ensemble and calibrated forecasts. Lower values indicate better calibration.

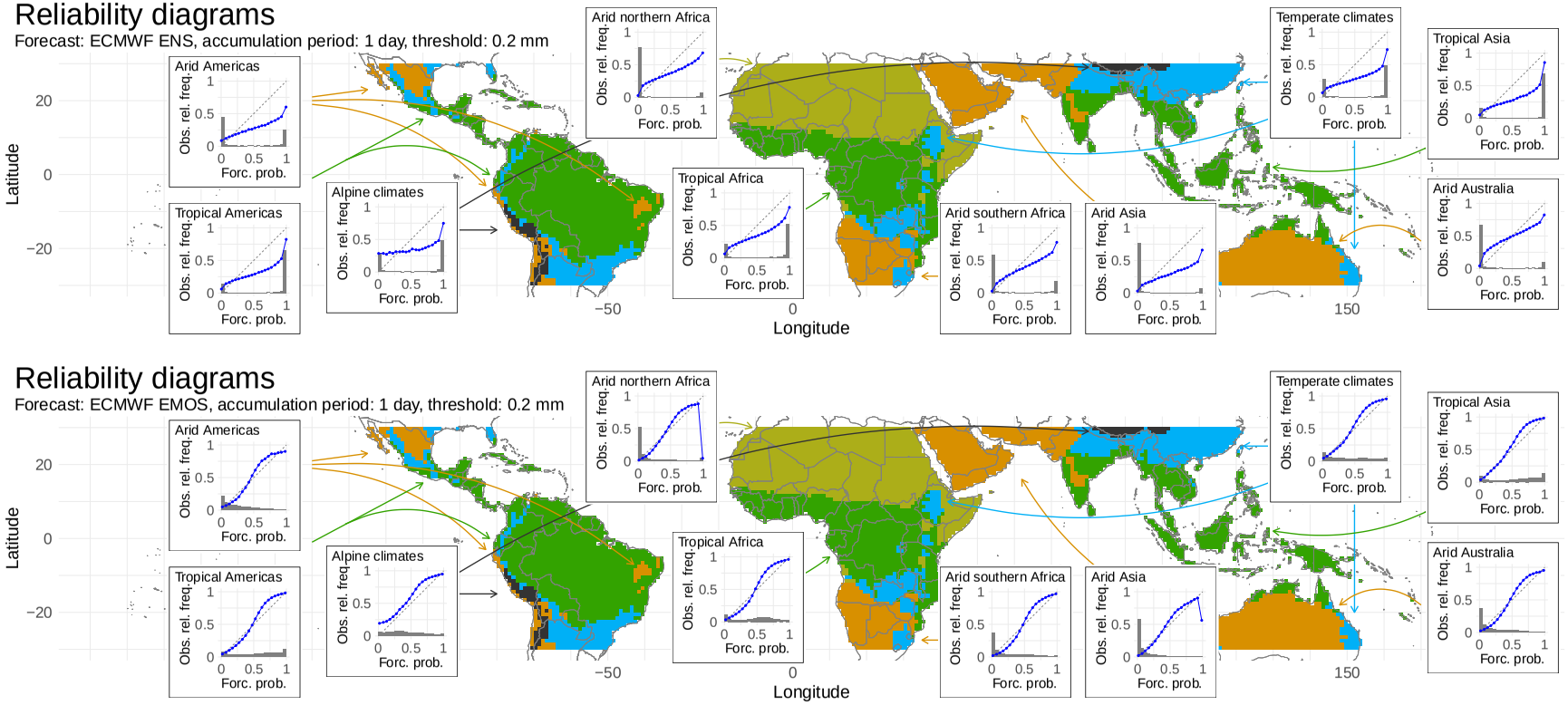


Figure 6.3: Reliability of ECMWF raw (top) and postprocessed (bottom) ensemble forecasts for the occurrence of precipitation within 1 day.

6.3.2 Skill of the ECMWF ensemble

Figure 6.4 displays the mean CRPS skill of raw and postprocessed ECMWF ensemble forecasts for 1-day accumulated precipitation and the period 2009–2017 relative to EPC. Over land, the ECMWF raw ensemble system issues skillful forecasts for some regions such as India, Australia or eastern Brasil, but struggles in complex terrain such as in the Himalayas or the Andes as well as in many places in tropical and northern arid Africa. Over oceans, the raw ensemble is skillful in many regions, but not for the oceanic deserts.

After postprocessing, the ECMWF ensemble forecast has almost everywhere neutral or positive skill. Regions with previously negative skill tend to have neutral skill, while regions with positive skill exhibit frequently a slightly higher positive skill. In particular, over complex terrain (e.g. Andes, Himalayas, mountain ranges of Papua New Guinea) and large parts of tropical and northern Africa, 1-day accumulated ECMWF ensemble forecasts exhibit only neutral skill, and have predictive performance equal only to EPC.

With longer accumulation periods, the distribution of ECMWF raw ensemble forecast skill changes. The top panel of Figure 6.5 displays CRPS skill for 5-day accumulated ECMWF raw ensemble precipitation forecasts. Compared to 1-day accumulations, raw ensemble forecast skill deteriorates slightly in regions where there is negative skill already for 1-day accumulation periods, and it increases slightly in many places that reveal positive skill for 1-day accumulation periods (e.g., Australia, arid Asia, arid southern Africa, eastern South America).

While CRPS and CRPSS allow to assess forecast quality with respect to the full probabilistic forecast, PoP as well as probabilities for the occurrence of accumulated precipitation above given thresholds are essential components of any precipitation forecast. The bottom panel of Figure 6.5 displays the BS skill of ECMWF raw ensemble forecasts for the occurrence of precipitation during the following 24 hours relative to EPC. ECMWF raw ensemble forecasts are skillful in few places only and have clear deficiencies in predicting the PoP over tropical oceans. Over land, the ECMWF raw ensemble is skillful in several regions such as southern Africa and southern Brasil, and quite skillful over Australia. Along the coast of East Africa and Brasil as well as in complex terrain, South East Asia, or tropical Africa, the ECMWF raw ensemble underperforms strongly when compared to EPC. Presumably, this is an indication of misrepresentation of convection and possibly its organization by the raw ECMWF EPS.

After postprocessing, the BS skill distribution for ECMWF PoP forecasts (not shown) is almost identical to the CRPSS skill map for 1-day accumulated precipitation displayed in Figure 6.4. At longer accumulation times, the negative skill of the ECMWF raw ensemble turns to neutral or only slightly negative skill in many parts of the oceans (not shown). Over land, the region of negative skill along the coast of East Africa expands inland, but reduces in tropical West Africa to a narrow band along the coast. In South America and Asia, the region of negative skill contracts and is almost exclusively confined to regions of complex terrain. ROC curves and Murphy diagrams (not shown) further support this analysis of

ECMWF PoP forecast quality.

6.3.3 Skill of ECMWF forecasts for extreme rainfall events

One important aspect of precipitation forecasts is their ability to predict extreme events such that precautionary action can be taken. Exemplarily, Webster et al. (2011) report on extreme rainfall events in Pakistan in 2010 that were embedded in the Indian monsoon during a period of anomalous large-scale flow and predicted by the ECMWF with high probabilities 6–8 days ahead. However, not all extreme precipitation events are connected to relatively well-predictable and large-scale features, and it is unclear if and where models are able to predict extreme precipitation some time ahead. As noted by Lerch et al. (2017), the immanent problem in evaluating forecast quality for extreme events is that sampling uncertainty typically impedes our ability to analyze forecast skill. One potential remedy is to increase the number of events. We achieve this by considering carefully selected thresholds and by relying on a long time series of events. Here, we use 20 mm within 24 hours and 50 mm within 5 days as thresholds for the occurrence of extreme events, and display results only for continents and those gridboxes where the considered event occurs with a frequency of at least 1%, or about 33 events in 2009–2017. Figure 6.6 displays BS skill for raw and postprocessed ECMWF forecasts for both types of extreme events. While ECMWF raw ensemble forecasts for the occurrence of extreme precipitation are mostly skillful throughout Asia, Australia, and the Americas, a clear lack of skill can be observed for Africa west of the East African rift. After postprocessing, the skill for the latter region is neutral, and positive almost everywhere else. Highest skill is observed for eastern China and eastern Australia, where the average BS skill for the prediction of these extreme events is as high as 0.30. At 5 days and 50 mm, raw ECMWF ensemble skill decreases in several regions, in particular Africa west of the East African ridge, western South America, and over complex terrain in Asia. Postprocessed ECMWF ensemble skill is very similar to 1 day and 20 mm, but typically higher such as in India or Brasil. For Africa west of the East African rift, ECMWF postprocessed ensemble forecasts have equal predictive performance as EPC.

What are the reasons for these results, in particular for the fact that ECMWF raw and postprocessed forecasts have predictive performance equal only to EPC forecasts for occurrence and amount of precipitation and extreme rainfall events in several tropical regions?

For the oceanic deserts, raw ensemble forecast skill is strongly negative. As TRMM is known to have a comparatively large dry bias in these regions (Huffman et al., 2007), TRMM-based EPC forecasts have a clear advantage over NWP raw ensemble forecasts and the skill of the latter is presumably assessed worse than it actually is.

Over complex terrain as in the Himalayas or Andes, TRMM performs relatively poor in the detection of precipitation (Barros et al., 2006; Hirpa et al., 2010; Maggioni et al., 2016). While observational deficiencies are one likely reason for

the lack of calibration and negative skill of both raw NWP ensembles in these regions, another is the insufficient representation of complex terrain in NWP models. Based on current resolutions, they can not fully resolve the orography which often results in lower predictive performance as analyzed, e.g., by Richard et al. (2007) for the European Alps.

For most of the remaining regions with negative skill of raw and neutral skill of postprocessed NWP forecasts and in particular for tropical and northern arid Africa, we suspect convective parameterization to be a major cause. In these regions, a special type of (ice scattering) MCSs account for a majority of rainfall as displayed in panels d) and f) of Figure 11 in Nesbitt et al. (2006). Recent studies suggest that convective parameterization often can not represent the high degree of convective organization in MCSs. This is a major impediment for their realistic representation in NWP models and leads to forecasts with too much light and too little intense rainfall overall (Stephens et al., 2010; Marsham et al., 2013; Pearson et al., 2014; Birch et al., 2014; Pantillon et al., 2015).

6.3.4 Comparison to the MSC ensemble

After assessing the skill of ECMWF raw and postprocessed forecasts for rainfall occurrence, rainfall amount, and extreme rainfall events, we now compare these results to those for the MSC ensemble and analyze differences and similarities.

Figure 6.7 displays calibration of the MSC ensemble in the same way as Figure 6.2, but with the maximal height of the PIT histograms in the top panel now being reduced to six instead of ten. While the MSC raw ensemble is also not calibrated, it is far better calibrated than the ECMWF raw ensemble, though slightly right skewed. The geographical distribution of the calibration of the MSC ensemble in the bottom panel reveals good calibration over large parts of the Indian and western Pacific oceans as well as in tropical Africa and northwestern South America. The reliability of the MSC raw ensemble is displayed in the top panel of Figure 6.8 and reveals that the MSC raw ensemble is clearly more reliable than the ECMWF raw ensemble for most regions. However, it also struggles to issue reliable forecasts for the occurrence of precipitation in alpine climates.

The spatial distribution of MSC raw ensemble BS skill is displayed in the bottom panel of Figure 6.8 and reveals skill for most regions, except oceanic deserts, parts of arid northern Africa and the Arabian peninsula, and in complex terrain. After postprocessing, the MSC ensemble is calibrated and as reliable as the ECMWF postprocessed forecast, but with a typically slightly lower resolution (not shown). The spatial distribution of CRPS skill for raw and postprocessed MSC ensemble forecasts for 1-day accumulated precipitation as displayed in Figure 6.9 is similar to that of the ECMWF ensemble. However, clear differences can be observed for tropical Africa where the MSC raw ensemble has neutral instead of negative skill, South America, where the negative skill of the MSC raw ensemble is restricted to the Andes region, and arid northern Africa, where the MSC ensemble performs worse than the ECMWF raw ensemble and EPC. The better skill in the MSC raw ensemble for tropical Africa and South America, however,

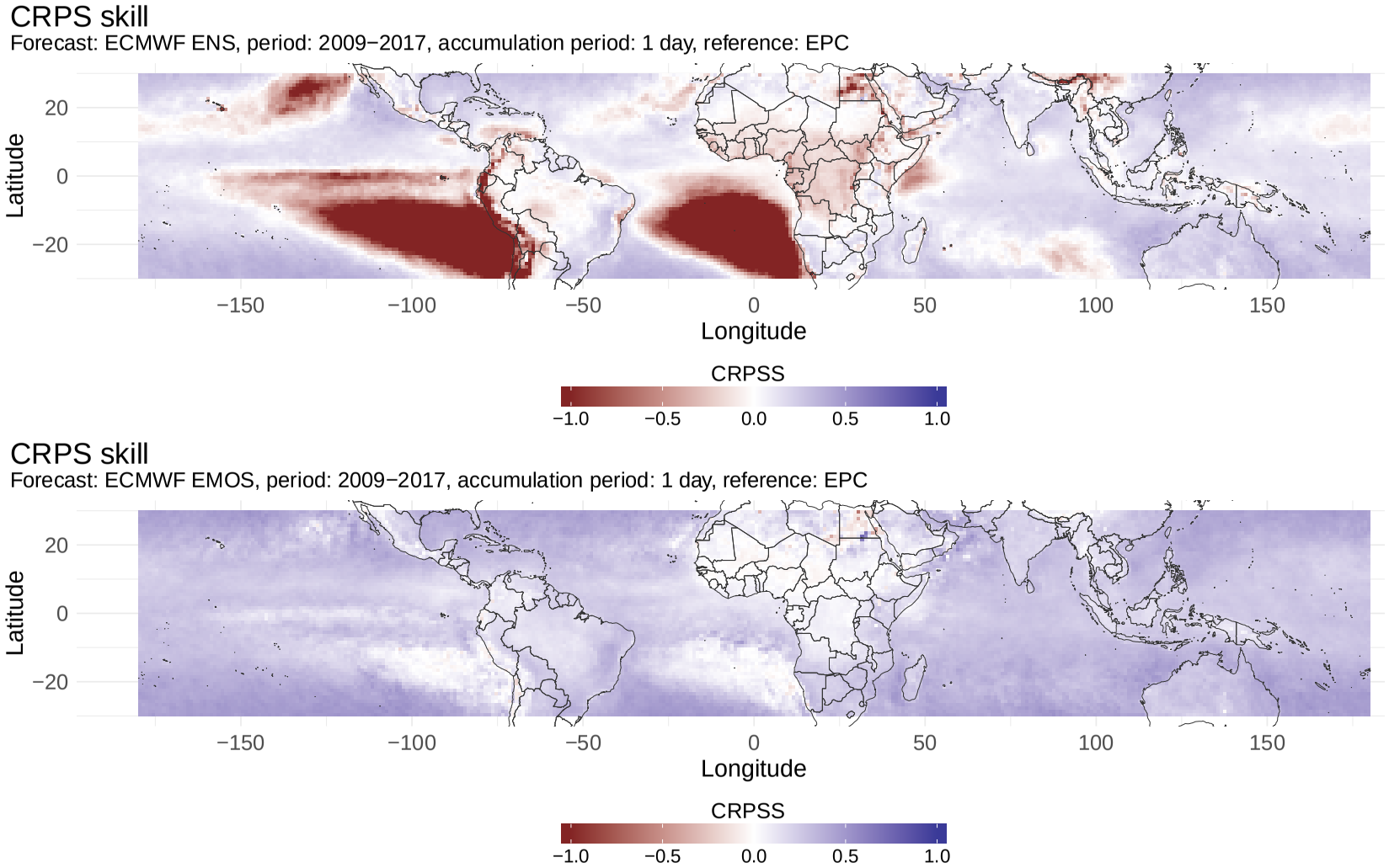


Figure 6.4: CRPS skill of ECMWF forecasts for 1-day accumulated precipitation. Displayed is the average CRPS skill of raw (top) and postprocessed (bottom) ECMWF ensemble forecasts for 1-day accumulated precipitation in 2009–2017 relative to EPC.

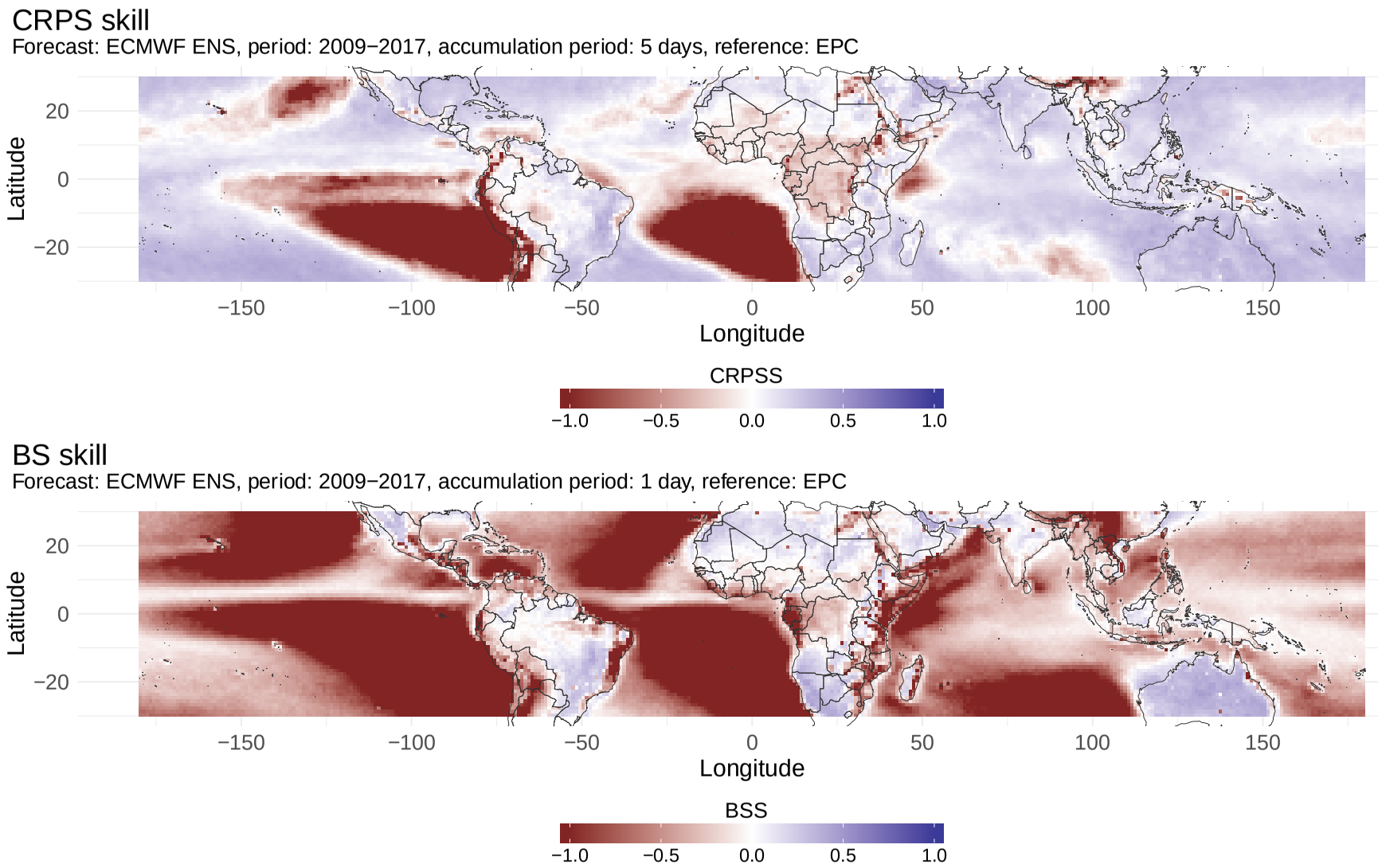


Figure 6.5: CRPS skill of ECMWF raw ensemble forecasts for 5-day (top) and BS skill of ECMWF raw ensemble forecasts for 1-day (bottom) accumulated precipitation. A threshold of 0.2 mm is used to determine the occurrence of precipitation.

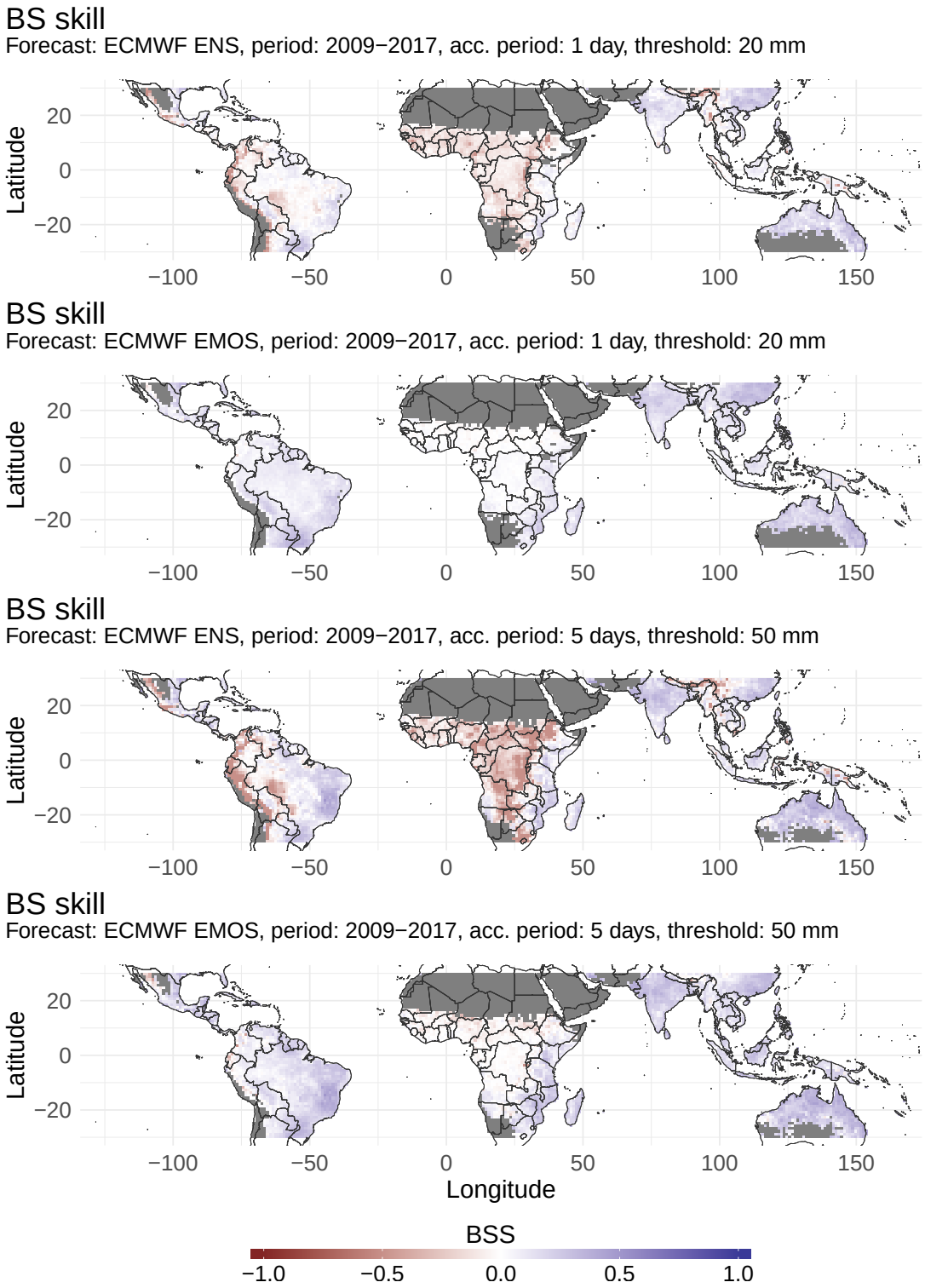


Figure 6.6: Predictability of extreme precipitation. BS skill for ECMWF raw and postprocessed ensemble forecasts for the exceedance of 20 mm within one day (top panels) and the exceedance of 50 mm within 5 days (bottom panels). Displayed is skill only over land and where the considered events has an occurrence frequency above 1%. The legend applies to all panels and gridboxes over land where the considered event has an occurrence frequency of less than 1% are gray.

does not yield better postprocessed predictions when compared to ECMWF.

This suggests that while the MSC raw ensemble is better calibrated and more reliable than the ECMWF raw ensemble in these two regions, it does not contain more predictive information than the latter. In particular, the higher BS and CRPS skill of the MSC raw ensemble compared to the ECMWF raw ensemble might only stem from better calibration and reliability. MSC raw and postprocessed ensemble forecasts for extreme rainfall are slightly worse than their ECMWF counterparts and have a very similar spatial distribution (not shown).

6.3.5 Improvement of ensemble forecasts from 2009 to 2017

In previous subsections, the ability of ECMWF and MSC raw and postprocessed ensemble forecasts to predict rainfall amount, occurrence, and extreme events was assessed with respect to the regional and spatial distribution based on the mean skill across 2009–2017. Due to the availability of verification data for the recent nine years, it is also possible to assess the change in skill over this period and thus the success in improving precipitation forecasts for the tropics over the last decade.

Figure 6.10 displays the temporal evolution of CRPS skill for raw and postprocessed ECMWF forecasts for 1-day accumulated precipitation in each Köppen-Geiger climate over the period 2009–2017. In 2009, ECMWF raw ensemble forecasts for all regions except for arid Asia and arid Australia have negative skill, in particular, forecasts for arid Americas, arid northern and tropical Africa, and for alpine climates. From 2009 to 2010, forecast skill increases strongly in most climates and becomes mostly positive. For some regions, the increase in skill continues until 2011. After 2011, no clear improvement in CRPS skill is detectable for most regions. We hypothesize that the increase in skill for ECMWF ensemble forecasts from 2009 to 2010 is at least partly related to the increase in horizontal resolution for all members as introduced on January 26, 2010, when the HRES run changed to a horizontal resolution of 16 km from previously 25 km, and the CNT and ENS runs to 32 km instead of 50 km (Miller et al., 2010).

As statistical postprocessing corrects for systematic forecast errors, postprocessed forecasts have neutral or positive skill in all Köppen-Geiger climates. The increase in skill gained by postprocessing is largest for alpine climates and arid Americas. Arid northern and tropical Africa reveal similar ECMWF raw ensemble skill as arid Americas, but much smaller ECMWF postprocessed ensemble forecast skill. Presumably, this indicates that forecasts for arid Americas contain more predictive information than for tropical and arid northern Africa, although the raw ensemble forecasts for the three regions have similar levels of miscalibration. With less convective organization in arid Americas (see, e.g., Nesbitt et al., 2006), this is a further indication of deficiencies of NWP ensembles in the representation of strongly organized convective systems.

Over the period 2009–2017, most regions reveal no or only slight increase in skill for postprocessed ECMWF ensemble forecasts. For most arid regions, any larger changes in postprocessed skill are observed for 2009 to 2010 and skill stays

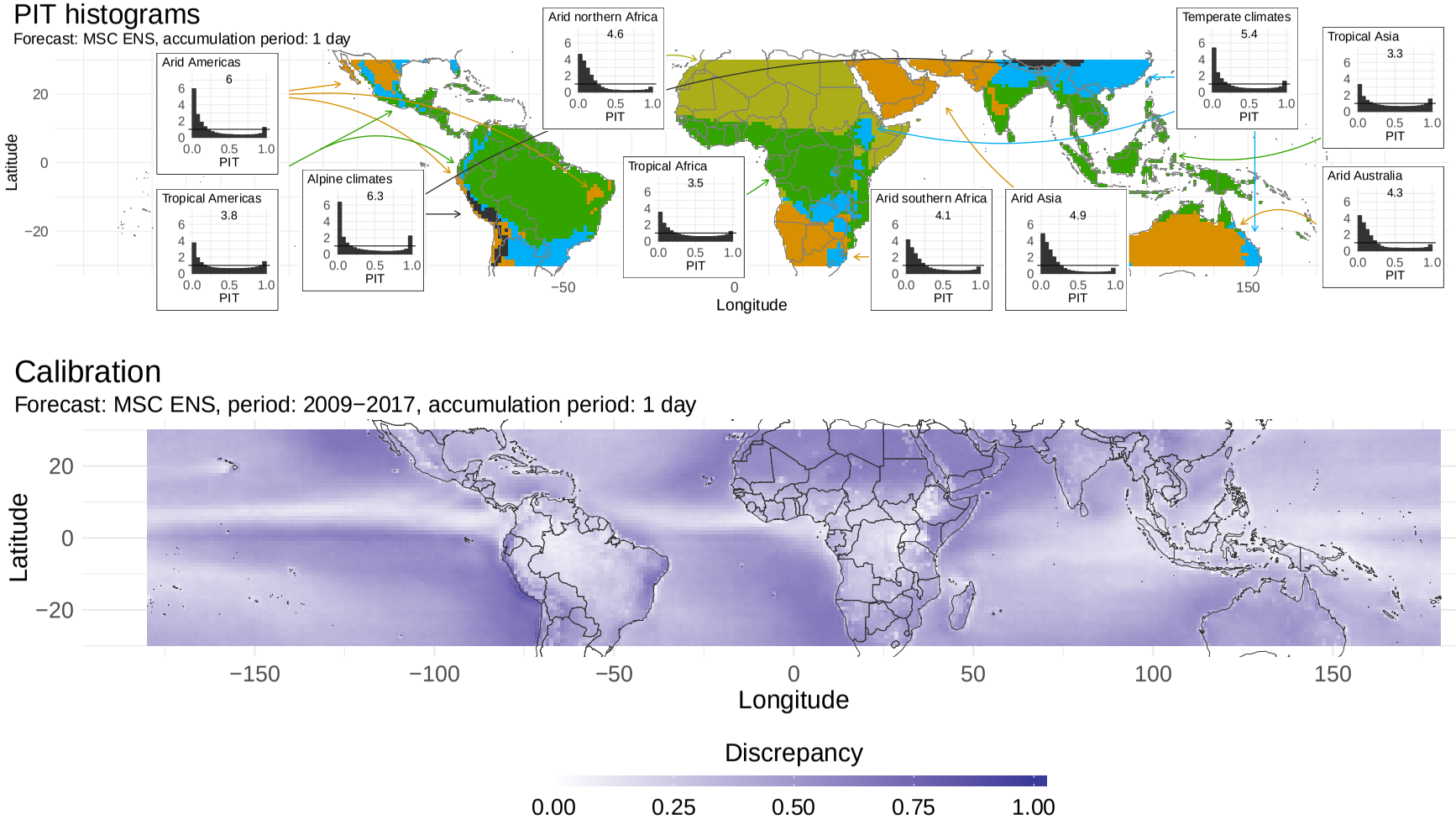


Figure 6.7: Calibration of MSC raw ensemble forecasts for 1-day accumulated precipitation. Panels are the same as in Figure 6.2, but for the MSC ensemble.

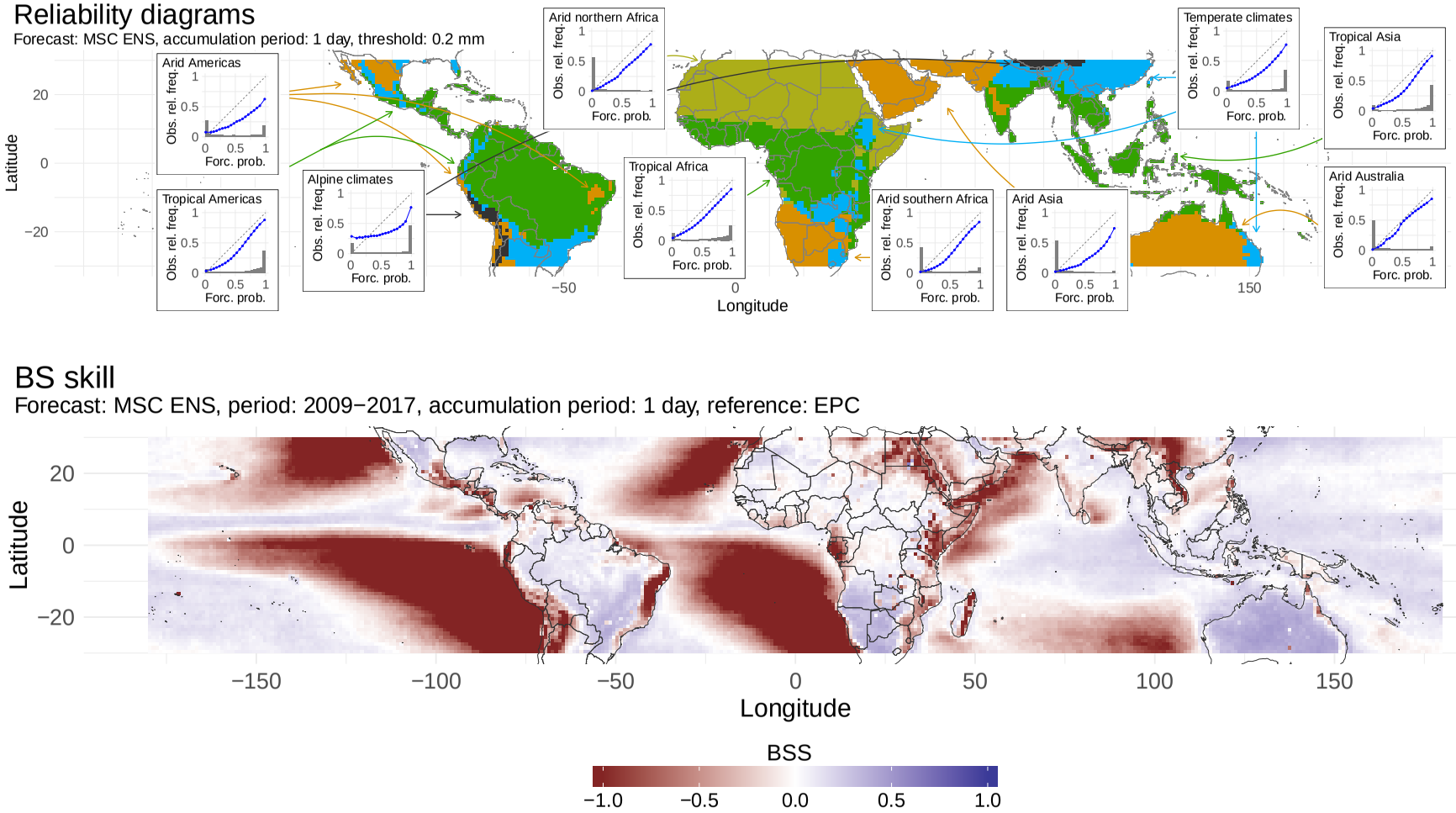


Figure 6.8: Reliability (top) and BS skill (bottom) of raw MSC forecasts for 1-day accumulated precipitation.

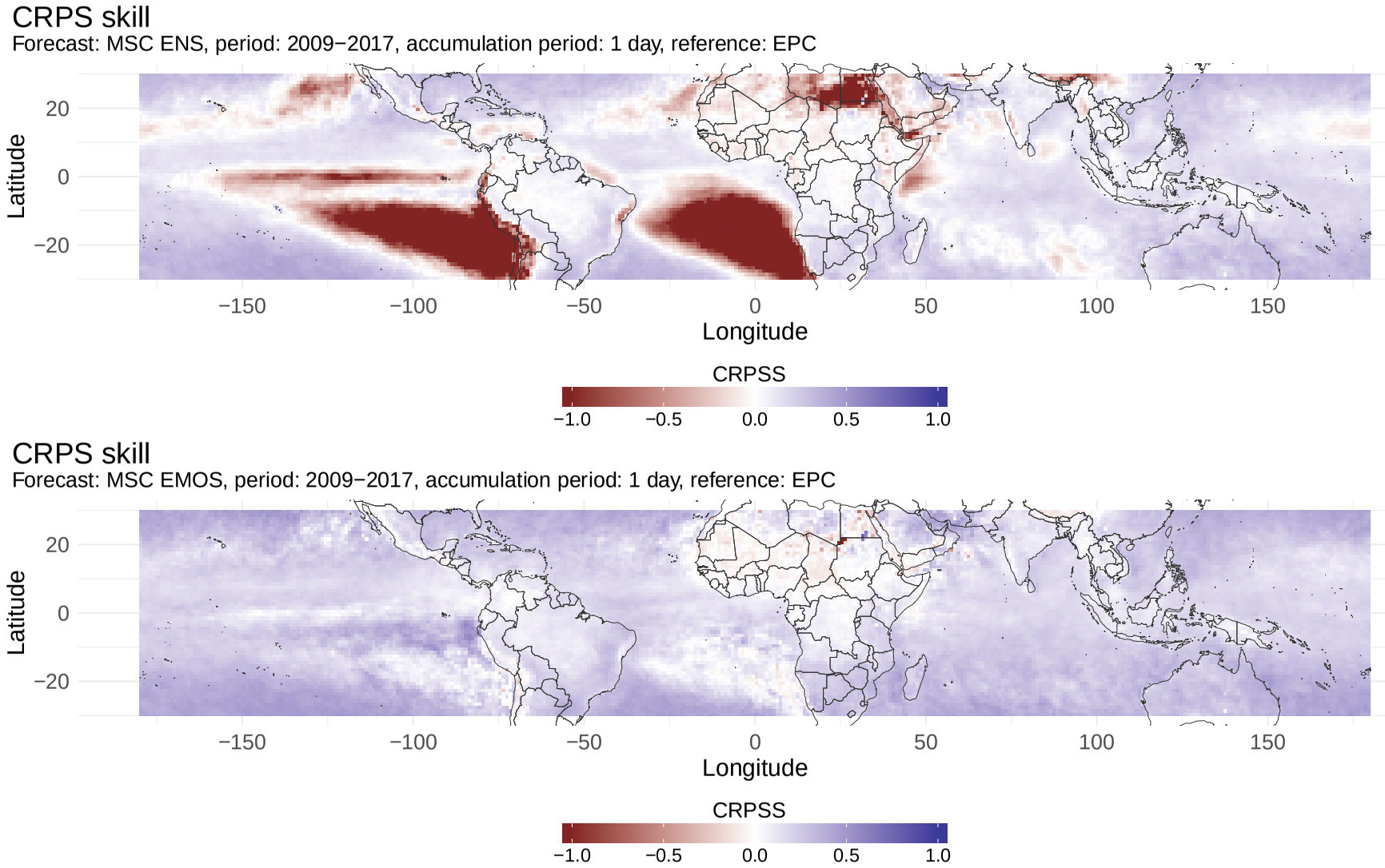


Figure 6.9: CRPS skill of MSC raw and postprocessed ensemble forecasts for 1-day accumulated precipitation.

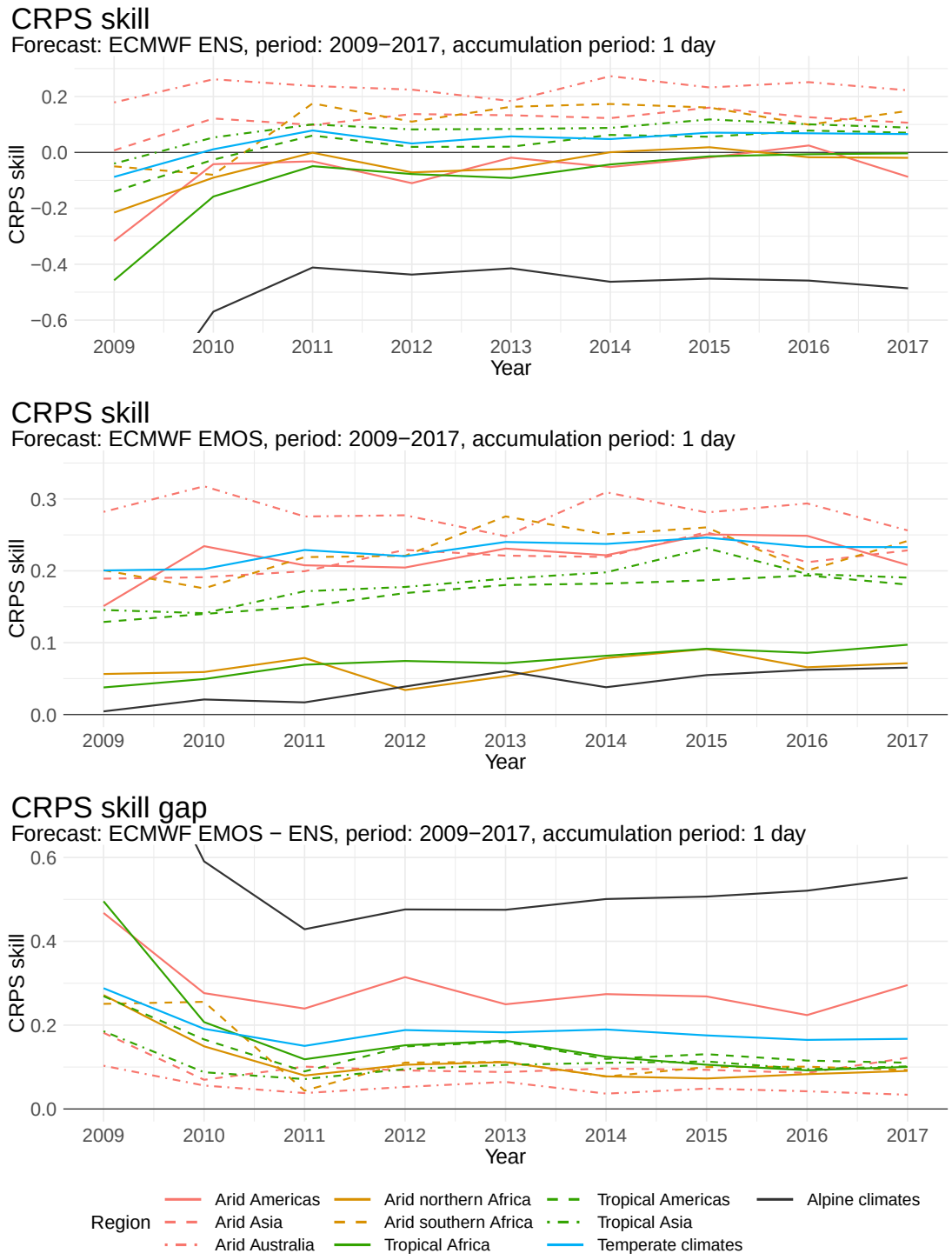


Figure 6.10: Improvement of forecast skill over 2009–2017. CRPS skill for raw (top) and postprocessed (middle) ECWMF forecasts for 1-day accumulated precipitation. The bottom panel displays the temporal evolution of the gap in skill between postprocessed and raw ECWMF forecasts for 1-day accumulated precipitation.

roughly constant afterwards. For the three tropical regions and tropical Africa in particular, postprocessed forecast skill continuously increases over 2009–2017 and yields an increase on the order of 5% over the nine years. Similarly, forecast skill in the alpine climates increases by about 6%, starting at almost no skill in 2009.

An interesting question regarding the temporal evolution of raw and postprocessed skill is the evolution of the skill gap. Hemri et al. (2014) investigate ECMWF forecasts of temperature and 1-day accumulated precipitation in this regard, verified against station observations, and find a constant improvement by postprocessing. The lower panel of Figure 6.10 displays the temporal evolution of the CRPSS skill gap between raw and postprocessed ECMWF forecasts. It shows a clear narrowing of the skill gap in all regions from 2009 to 2011, where it decreases for the majority of regions from a level between 10% and 30% down to 5%–15%. After 2011, however, the gap in skill remains about constant in most regions and increases even slightly for alpine climates.

For the MSC model, our results are similar to those for the ECMWF, and we briefly summarize the differences. Already in 2009, MSC raw ensemble forecasts have neutral or slightly positive CRPS skill in all regions except alpine climates and are slightly more skillful than the ECMWF raw ensemble in most regions (not shown). Postprocessing improves MSC forecast skill, but MSC postprocessed forecasts are less skillful than postprocessed ECMWF forecasts in all regions (not shown).

6.4 Discussion

For the tropics and ten continental, tropical Köppen-Geiger climates, the quality of raw and postprocessed accumulated precipitation forecasts from two leading operational EPSs has been assessed for several years and accumulation periods. In particular, we have examined the ability of ECMWF and MSC ensemble forecasts to issue predictions for rainfall amount, occurrence, and extremes relative to the climatological reference EPC. Both raw ensembles exhibit clear calibration problems and are overconfident and unreliable in the prediction of the PoP. For several Köppen-Geiger climates such as arid Australia or arid southern Africa, both raw ensembles are skillful, while they have at best neutral skill for tropical and northern arid Africa and in alpine climates.

After correcting for systematic forecast errors by statistical postprocessing, forecasts for amount and occurrence of precipitation are skillful in several climates. In tropical and northern arid Africa and alpine climates, however, even postprocessed NWP forecasts have predictive performance equal only to EPC. This suggests that for these regions NWP ensemble forecasts can not provide more insights about precipitation events in the near future than a climatological forecast, even though the former have access to recent information on the state of the atmosphere. As thresholds for extreme events, we considered 20 mm within one day and 50 mm within five days and found skill in ECMWF and MSC

ensemble forecasts for the prediction of these events in most climatic regions. Exceptions are as before tropical and northern arid Africa and alpine climates.

We suspect three main problems, namely convective parameterization, model resolution, and observational errors as major causes for the poor performance of raw and postprocessed NWP ensemble predictions in these regions. In alpine climates with complex terrain and steep orography, TRMM observations have known deficiencies in the detection of rainfall (e.g., Barros et al., 2006; Hirpa et al., 2010), but also NWP models are known to have lower predictive performance (Richard et al., 2007). As such, raw NWP ensemble forecasts for alpine climates are likely assessed worse than they actually are, but, presumably, they are also not skillful.

In almost all other regions with negative skill in raw and neutral skill in postprocessed ensemble forecasts, MCSs account for a major proportion of total rainfall. Convective parameterization often struggles to represent the high degree of organization in MCSs and leads to too many and too weak rain events overall (see, e.g., Marsham et al., 2013). For tropical and northern arid Africa, one can also suspect that the lack of ground-based observations and radiosondes is an obstacle for skillful forecasts. For West Africa, Agustí-Panareda et al. (2010) show that the predictive information gained by assimilating numerous radiosonde soundings during the AMMA field campaign was typically lost in less than 24 hours. Thus, improving forecast quality for this region can not be achieved by additional observations only, but requires better NWP models and process understanding. As such, the scarcity of (good) observational data is at least an indirect cause of the poor predictive skill of NWP ensembles in tropical and northern arid Africa.

Over the investigation period 2009–2017, ECMWF raw ensemble skill increases strongly between 2009 to 2011 and only marginally afterwards. Raw MSC forecasts are more skillful than their ECMWF counterparts, but their skill improves only slightly across 2009–2017. After statistical postprocessing, the skill of both NWP models reveals only small improvements over 2009–2017. Most improvements are confined to the moist tropical Köppen-Geiger climates and presumably mirror the slow improvement from model physics development. While sudden increases in forecast skill for accumulated precipitation should not be expected, it is disconcerting that for most arid climates in the tropics no improvement in forecast skill, neither for raw nor postprocessed forecasts, can be observed for a period of almost one decade. This standstill in forecast improvement for accumulated precipitation forecasts requires further attention and investigation.

7 Statistical forecasts for the occurrence of precipitation in northern tropical Africa

In a comprehensive study, we have investigated in Chapter 5 the predictive skill of nine global NWP ensembles for accumulated precipitation in northern tropical Africa. Raw ensemble forecasts are uncalibrated and unreliable and underperform in the prediction of occurrence and amount of precipitation when compared to EPC. This assessment is robust and holds for all regions, accumulation periods, monsoon season, and TIGGE sub-ensembles. After statistical postprocessing, forecasts are calibrated and reliable, but only have equal predictive performance when compared to EPC.

In Chapter 6, we have extended this assessment to the tropics between 30°S and 30°N and examined forecast quality for ten continental Köppen-Geiger climates. Despite a lack of calibration and reliability, raw ECMWF and MSC ensemble forecasts are skillful for a majority of climatic regions. Postprocessing improved the predictive performance of both models and lead to skillful forecasts in most climates. However, for northern arid and tropical Africa as well as alpine climates, we found no skill of raw and postprocessed ensemble forecasts for the prediction of rainfall occurrence, amount, or extremes at accumulation periods of 1 to 5 days.

While the results for these regions are disappointing and require further investigation, they call in addition for alternative approaches for the prediction of occurrence and amount of precipitation. In particular, one can ask whether it is possible to construct probabilistic precipitation forecasts that rely on a climatological baseline, but obtain higher sharpness and resolution by involving recent observations of meaningful atmospheric variables or events. In this chapter, we investigate one alternative approach and present results of our investigations. In Section 7.1 the spatio-temporal correlation of rainfall and its occurrence is analyzed. In Section 7.2, we construct a statistical forecast and evaluate its forecast quality for northern tropical Africa and 1998–2014. Section 7.3 concludes.

7.1 Spatio-temporal correlation of precipitation

As any probabilistic precipitation forecast consists of and can be split into the PoP and a probability distribution for the amount of precipitation, we restrict the prediction of precipitation in the following to the PoP. In parts of northern

tropical Africa, MCSs account for the majority of rainfall and are frequently coupled with AEWs. While MCSs typically propagate for one or two days only, the coupling with AEWs allows for propagation of MCSs properties beyond the decay of the actual MCS. We assume that the presence an MCS increases the PoP downstream of its current location for the near future, while its absence decreases the PoP. To evaluate the validity of this assumption, we exemplarily analyze the relationship between 1-day accumulated precipitation at Niamey, the capital of Niger, and lagged observations of 1-day accumulated rainfall for the months July–September 1998–2013 using Spearman's rank correlation. In the frequently occurring case of ties in 1-day accumulated precipitation observations, we assign to the subset of these observations the average rank of the subset, and compute Spearman's correlation as Pearson correlation of the ranks. We base our investigation on TRMM observations to obtain a full spatial coverage and use a spatial aggregation of $1^\circ \times 1^\circ$.[1]

For lags of one and two days, Figure 7.1 displays the locations and correlations for those gridboxes that have correlation coefficients higher than the 0.99 or lower than the 0.01 quantile of all correlation coefficients in our study regions. We find the highest positive correlation coefficients east of Niamey, well clustered, and at a distance that is slightly larger than the average distance AEWs travel in one or two days, and slightly lower than the average distance MCSs travel in one or two days. This coincides with our assumption that information on recent rainfall events propagates with MCSs and AEWs. The interpretation of the negative correlation is less clear and we suspect that precipitation events southwest and southeast of Niamey are indicators for the advance or retreat of the West African monsoon that modulates amount and occurrence of precipitation at Niamey (see, e.g., the EPC forecast in Figure 7.3).

Figure 7.2 focuses on the modulation of the PoP at Niamey conditional on the accumulated precipitation amounts at the highest positively correlated locations at lags of one and two days. We classify these precipitation amounts into no, light, and strong precipitation where the latter two correspond to precipitation amounts below and above the climatological median of all positive precipitation amounts at the respective location in the considered period. The top left panel displays the PoP at Niamey conditional on the categorical precipitation event at lag one. From an average climatological PoP of 0.44 for the period July–September, the PoP reduces to 0.35 if no precipitation was observed at lag one in the corresponding location, and increases to 0.62 if strong precipitation occurred there at lag one. For the categorical observations at lag two, similar findings hold though the deviations from the climatological PoP are slightly smaller. Considering both observations jointly reveals even stronger modulations of the PoP as displayed in the bottom left panel. If at both lags no precipitation was observed at the corresponding locations, then the PoP at Niamey is as low as 0.31, while it is 0.76 if at lags of one and two days strong precipitation was observed.

[1] We have also studied the spatio-temporal correlation of precipitation at the native resolution of TRMM of $0.25^\circ \times 0.25^\circ$ and obtained similar results.

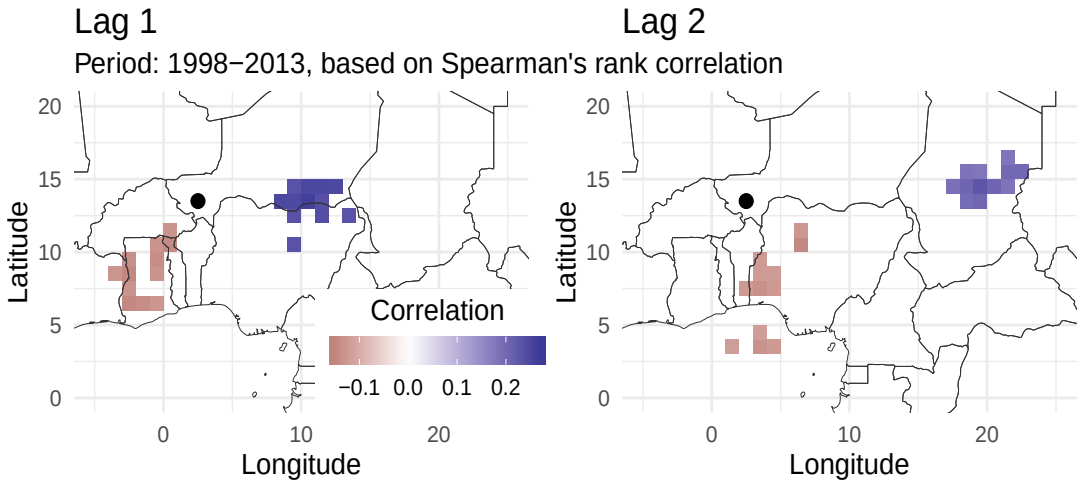


Figure 7.1: Spatio-temporal correlation of precipitation. Displayed are the highest 1% positive (blue) and negative (red) correlations between 1-day accumulated precipitation at Niamey (•) and 1-day accumulated precipitation in northern tropical Africa at lags of one (left) and two (right) days, based on Spearman's rank correlation and the period July–September 1998–2013.

7.2 Statistical forecasts for the occurrence of precipitation

While these results allow to build a forecast for the PoP at Niamey based on the histogram in Figure 7.2, such a forecast is suboptimal as it necessitates a discretization of the real-valued amount of precipitation. Logistic regression models do not have this problem and we rely for the prediction of the PoP p at Niamey on logistic regression forecasts of the form

$$\begin{aligned}\text{logit } p \,|\, o_{1+}, o_{2+}, o_{1-}, o_{2-}, d = \\ s(d) + a_{1+}\, f(o_{1+}) + a_{2+}\, f(o_{2+}) + a_{1-}\, f(o_{1-}) + a_{2-}\, f(o_{2-}),\end{aligned} \tag{7.1}$$

that have also been explored by Klar (2017). Here, $f(x) = \log(x + 0.001)$ is a transformation of the amount of precipitation and

$$s(d) = b_0 + b_1 \sin\left(\frac{2\pi d}{365}\right) + b_2 \cos\left(\frac{2\pi d}{365}\right) \tag{7.2}$$

a parametric periodic function that depends on the day of the year d only. The observations at lags of one and two days at the strongest positively and negatively correlated locations are denoted by o_{1+}, o_{2+}, o_{1-}, and o_{2-}, respectively. In the following, we explain the generation of such logistic regression based forecasts and distinguish between verification and training data. When issuing predictions for

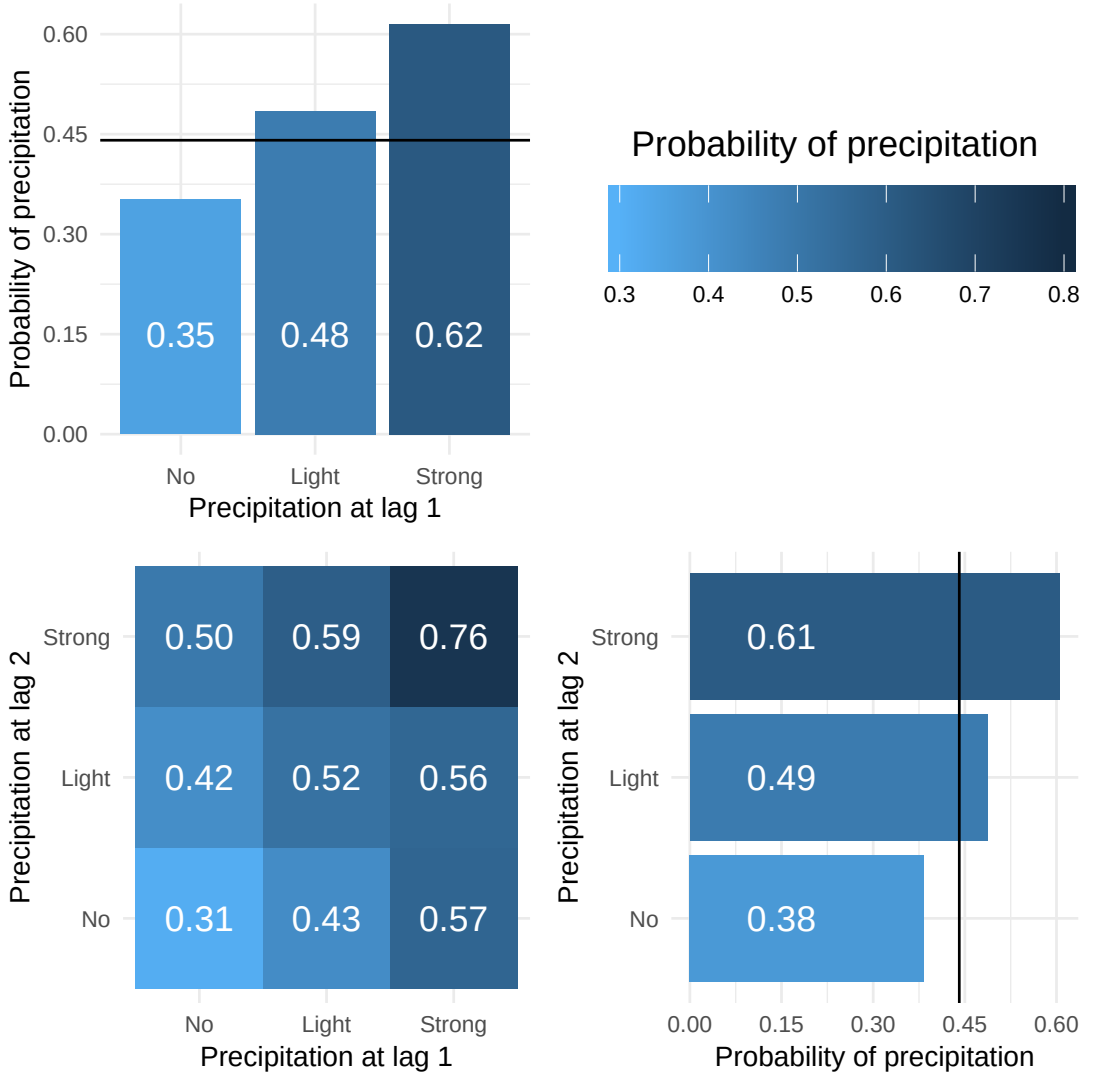


Figure 7.2: Modulation of the PoP at Niamey by 1- and 2-day lagged precipitation observations. The top left and bottom right panels display the modulation of the PoP at Niamey conditional on 1- or 2-day lagged categorized precipitation for the period July–September 1998–2013. The black line indicates the climatological PoP of 0.44 at Niamey for the same period. The bottom left panel displays the modulation of the PoP at Niamey when both 1- and 2-day lagged categorized precipitation is known, and the top left panel displays the color coding for the PoP.

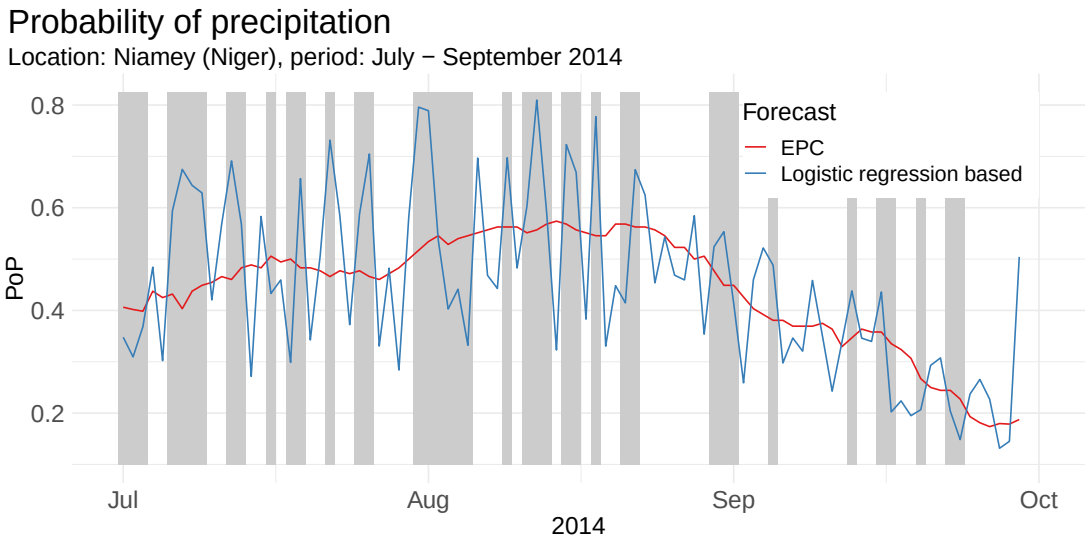


Figure 7.3: EPC (red) and logistic regression based (blue) forecasts for the occurrence of precipitation at Niamey during July–September 2014, and the actual occurrence of precipitation indicated by gray shading.

a given year in 1998–2014, the verification data contain only observations from July–September of the considered year, and the training data all observations from July–September 1998–2014 except for the considered year. The advantage of this approach is that estimation of the EPC and the logistic regression based forecast is performed on the same data set, which allows for meaningful comparisons. Using only the training data, we estimate Spearman's rank correlations, identify the locations with the highest positive and negative correlation coefficients at lags of one and two days, and estimate the parameters of the logistic regression model in (7.1). The actual PoP forecast for each day of the considered year is then based on $o_{1+}, \ldots, o_{2-}, d$ of the verification data and the parameter estimates.

Exemplarily, Figure 7.3 displays logistic regression based and EPC forecasts with a window length of ± 2 days for the occurrence of precipitation at Niamey in July–September 2014 and the verifying observations. Logistic regression based PoP forecasts clearly follow the increase in the climatological PoP during July and August as well as the decrease in the climatological PoP during September that is associated with the advance and retreat of the West African monsoon. Additionally, they deviate quite frequently by ± 0.1 and more from the climatological PoP and seem, in particular for July and August 2014, to outperform EPC in the prediction of the occurrence of precipitation at Niamey.

For a comparative assessment of reliability and resolution of the logistic regression based forecast, Figure 7.4 displays reliability diagrams for both forecasts. As a reliability diagram based on the period July–September of only one year does not provide meaningful insights, we rely on the full period July–September 1998–2014 instead and generate logistic regression based forecasts in the same fashion

as described earlier. This yields a meaningful cross-verification of the quality of the logistic regression based forecast. As shown in Figure 7.4, logistic regression forecasts have higher resolution and equal reliability as EPC.

We extend this approach to all gridboxes in our study region and assess the quality of the logistic regression based forecast relative to EPC by the BS skill. Figure 7.5 displays the spatial distribution of the mean BS skill of the logistic regression based forecast for the period July–September 1998–2014 relative to EPC. Almost everywhere in northern tropical Africa, logistic regression based forecasts have positive skill and outperform EPC and current NWP ensemble forecasts in the prediction of PoP. The Sahel, West Africa, and Ethiopia largely exhibit clear positive skill, while South Sudan and Central African Republic typically reveal BS skill between 0.0 and 0.1. The spatially and temporally averaged BS skill is slightly above 0.20.

7.3 Discussion

In this chapter, we have examined whether it is possible to construct skillful statistical forecasts for precipitation in northern tropical Africa by using information about recent rainfall events. Based on meteorological knowledge about propagation, organization, and coupling of convective systems in this region, we assumed that MCSs modulate PoP downstream of their current location. Exemplarily, we evaluated the validity of this assumption at Niamey, and found a modulation of the PoP by recent rainfall events of up to $\pm 20\%$. Using precipitation observations of the last two days, we constructed logistic regression based PoP forecasts for Niamey. These forecasts are reliable and have a higher resolution than EPC. In an extension to northern tropical Africa and the period 1998–2014, we found clear improvements in predictive performance by logistic regression based PoP forecasts with a spatially and temporally averaged BS skill of slightly above 0.20. Hence, statistical PoP forecasts based on recent information about the state of the atmosphere are an attractive alternative for forecasting precipitation and should be further investigated.

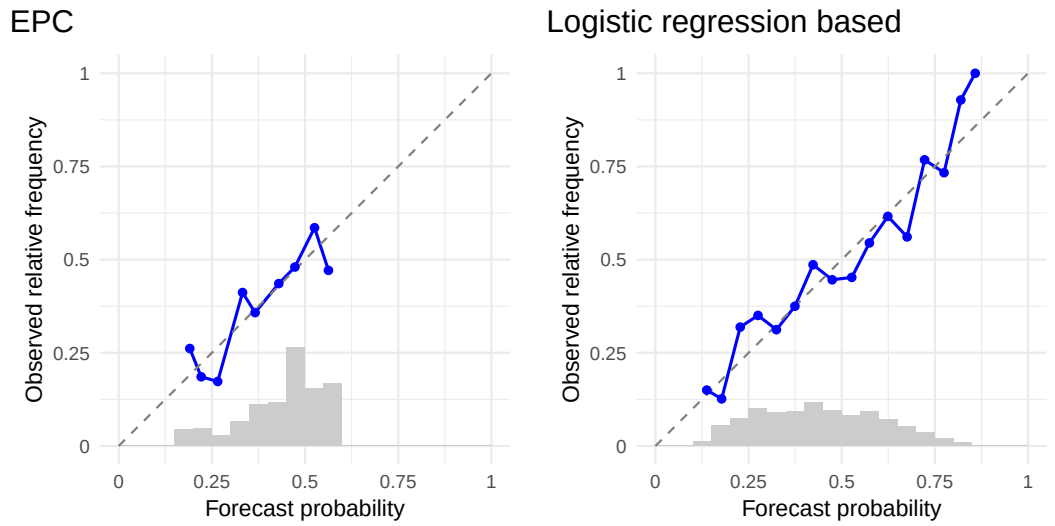


Figure 7.4: Comparative assessment of reliability and resolution of logistic regression based PoP forecasts. Displayed are reliability diagrams for EPC (left) and logistic regression based (right) PoP forecasts for Niamey and the period 1998–2014.

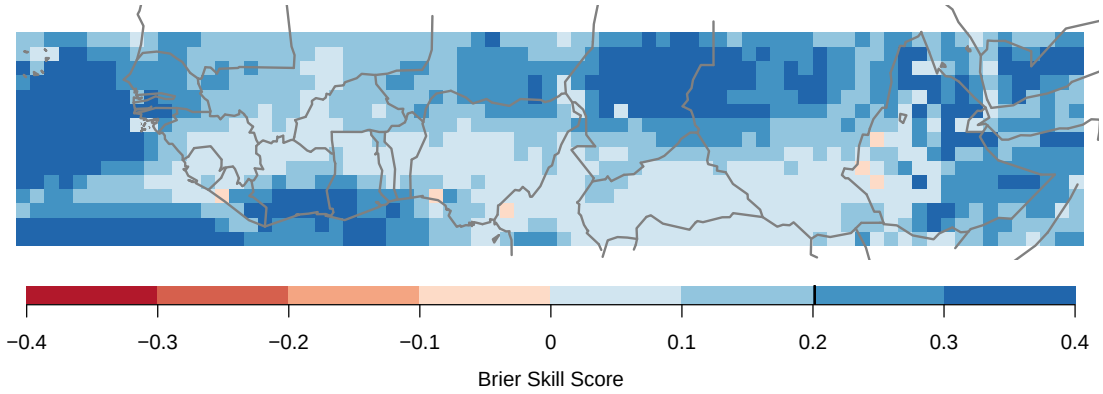


Figure 7.5: Spatial map of mean BS skill for logistic regression based PoP forecasts relative to EPC with a window length of ±2 days and for the period July–September 1998–2014. Temporally and spatially averaged BS skill of logistic regression based PoP predictions is slightly higher than 0.20.

8 | Conclusion

This thesis has demonstrated from theoretical and applied perspectives how the performance of probabilistic predictions can and should be assessed. The following concluding remarks start with a theoretical perspective before turning to an applied one, thereby summarizing and discussing our key results.

ROC curves are frequently used tools for the assessment of potential predictive skill for binary events as they are easy to interpret and have further desirable properties. We proved a near-equivalence between ROC curves and CDFs on the unit interval and elucidated the essential constraint of concavity that guarantees nondecreasing conditional event probabilities. For the classical binormal model and its generalization, we showed in Theorem 3.14 that ROC curves based on the location-scale approach are necessarily non-concave whenever the variances of the two conditional distributions are different. This strongly inhibits the applicability of the current approach and calls for alternatives.

As suggested by the near-equivalence of ROC curves and CDFs on the unit interval, we propose to model ROC curves by beta distributions and illustrate that beta ROC curves are, especially under the constraint of concavity, more flexible that binormal ones. For the estimation of model parameters, we rely on MD estimation and derive the asymptotic distribution of the MD estimator. Based on the asymptotic normality of the MD estimator, we construct tests for goodness-of-fit, equality of ROC curves, and equal predictive ability. Turning to empirical examples, we find empirical evidence for the increased flexibility of beta ROC curves when compared to binormal ones. These methods have been implemented and are currently prepared for release as freely available software package `betaROC` for the statistical programming language `R` (R Core Team, 2018).

NWP ensembles are one key application of probabilistic forecasting, and the prediction of accumulated precipitation is particularly challenging. In Chapter 5, we investigated the predictive skill of nine global NWP ensembles for 1–5 day accumulated precipitation in three regions in northern tropical Africa. For verification we relied on station and gridded satellite-based observations at different spatial aggregations and made assessments relative to a climatological reference forecast coined EPC. Raw ensemble forecasts are uncalibrated and unreliable and clearly less skillful in predicting occurrence and amount of precipitation than EPC, independently of region, accumulation time, monsoon season, and ensemble. Differences between raw ensembles and EPC are large and partly stem from poor predictions for low precipitation amounts. Statistical postprocessing by EMOS and BMA ensures calibration of ensemble forecasts, but very often at the cost of increased forecast uncertainty and lower resolution. Postprocessed forecasts are calibrated, reliable, and strongly improve on the raw ensembles, but are

typically not able to outperform EPC. This negative result is disappointing and unexpected. It suggests that current NWP ensembles are not able to translate any recent information on the state of the atmosphere into meaningful probabilistic statements regarding occurrence or amount of precipitation. We suspect convective parameterization to be a likely cause for these results and note that alternative approaches for the prediction of occurrence and amount of precipitation should be investigated.

Chapter 6 continued the assessment of the predictive performance of NWP ensemble forecasts for accumulated precipitation, extended it to the tropics between latitudes 30°S, and 30°N, and additionally assessed predictions for extreme rainfall events. We relied on satellite-based rainfall estimates for spatially consistent and complete observations and verified ensemble predictions for accumulation periods of 1–5 days and the years 2009–2017. Raw ensemble forecasts are uncalibrated and unreliable, and despite these deficiencies slightly skillful for several climatic regions within the tropics. Statistical postprocessing yields calibrated ensemble forecasts that have higher skill than raw ensemble forecasts and are reasonably skillful in most regions. From 2009 to 2017, we find little to no improvements in postprocessed forecast skill for all arid climates. This is disconcerting and requires further investigation. In western and tropical Africa but also over complex terrain, even postprocessed NWP ensemble forecasts for amount and occurrence of precipitation as well as extreme rainfall event have only neutral skill. These results are in agreement with our findings in Chapter 5 for northern tropical Africa and suggest further regions where statistical forecasts can be an attractive alternative to NWP ensemble forecasts for the prediction of accumulated precipitation.

In Chapter 7, we investigated an alternative approach for the prediction of PoP in northern tropical Africa. Coinciding with meteorological knowledge about convective systems in this region, the PoP is strongly modulated by recent rainfall events. This allows to construct statistical forecasts for the PoP that are reliable and have higher resolution than EPC. Across northern tropical Africa and 1998–2014, logistic regression based forecasts yield an average improvement of 20% above EPC and NWP postprocessed forecasts, and are a particularly attractive alternative for forecasting the PoP. As our investigations in Chapter 7 close with the assessment of the predictive performance of logistic regression based PoP forecasts for northern tropical Africa, this leaves interesting scientific questions for further research. In particular, we have so far only specified the PoP and not provided probabilistic forecasts for the amount of precipitation. While one can combine logistic regression based forecasts for the PoP with a climatological distribution for the amount of precipitation, our results suggest that also the amount of precipitation is modulated by the amount of recent rainfall events. As triggering, growth, and propagation of convective systems depend on many variables, we assume that some of these influence the PoP or the amount of precipitation. Exemplarily, cold pools are known triggers for convective systems, mid-level shear is regarded an essential ingredient for convective organization, and the availability of moisture in the Sahel might limit the vigorousness and amount

of precipitation. Finally, it remains open if our approach can be successfully applied in other regions such as tropical Africa and alpine climates.

Bibliography

Agustí-Panareda, A., Beljaars, A., Cardinali, C., Genkova, I. and Thorncroft, C. (2010). Impacts of assimilating AMMA soundings on ECMWF analyses and forecasts. *Weather and Forecasting*, 25, 1142–1160.

Anderson, J. L. (1996). A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of Climate*, 9, 1518–1530.

Atger, F. (1999). The skill of ensemble prediction systems. *Monthly Weather Review*, 127, 1941–1953.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T. and Silvermann, E. (1955). An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, 26, 641–647.

Barros, A. P., Chiao, S., Lang, T. J., Burbank, D. and Putkonen, J. (2006). From weather to climate-seasonal and interannual variability of storms and implications for erosion processes in the Himalaya. *Special Papers-Geological Society of America*, 398, 17.

Bauer, P., Thorpe, A. and Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, 525, 47–55.

Bechtold, P., Semane, N., Lopez, P., Chaboureau, J.-P., Beljaars, A. and Bormann, N. (2014). Representing equilibrium and nonequilibrium convection in large-scale models. *Journal of the Atmospheric Sciences*, 71, 734–753.

Berrocal, V. J., Raftery, A. E. and Gneiting, T. (2007). Combining spatial statistical and ensemble information in probabilistic weather forecasts. *Monthly Weather Review*, 135, 1386–1402.

Birch, C. E., Parker, D. J., Marsham, J. H., Copsey, D. and Garcia-Carreras, L. (2014). A seamless assessment of the role of convection in the water cycle of the West African monsoon. *Journal of Geophysical Research: Atmospheres*, 119, 2890–2912.

Bjerknes, V. (1904). Das Problem der Wettervorhersage, betrachtet vom Standpunkte der Mechanik und der Physik. *Meteorologische Zeitschrift*, 21, 1–7.

Bougeault, P., Toth, Z., Bishop, C., Brown, B., Burridge, D., De Chen, H., Ebert, B., Fuentes, M., Hamill, T. M., Mylne, K., Nicolau, J., Paccagnella, T., Park, Y. Y., Parsons, D., Raoult, B., Schuster, D., Dias, P. S., Swinbank, R.,

Takeuchi, Y., Tennant, W., Wilson, L. and Worley, S. (2010). The THORPEX interactive grand global ensemble. *Bulletin of the American Meteorological Society*, 91, 1059–1072.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30, 1145–1159.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78, 1–3.

Bröcker, J. and Smith, L. A. (2007). Increasing the reliability of reliability diagrams. *Weather and Forecasting*, 22, 651–661.

Buizza, R., Barkmeijer, J., Palmer, T. N. and Richardson, D. S. (2000). Current status and future developments of the ECMWF ensemble prediction system. *Meteorological Applications*, 7, 163–175.

Buizza, R., Houtekamer, P. L., Toth, Z., Pellerin, G., Wei, M. and Zhu, Y. (2005). A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Monthly Weather Review*, 133, 1076–1097.

Davis, J., Knippertz, P. and Fink, A. H. (2013). The predictability of precipitation episodes during the West African dry season. *Quarterly Journal of the Royal Meteorological Society*, 139, 1047–1058.

Dawid, A. P. (1984). Present position and potential developments : Some personal views : Statistical theory : The prequential approach. *1Journal of the Royal Statistical Society. Series A: Statistics in Society*, 147, 278–292.

De Leeuw, J., Hornik, K. and Mair, P. (2009). Isotone optimization in R: Pool-Adjacent-Violators Algorithm (PAVA) and active set methods. *Journal of Statistical Software*, 32, 1–24.

DeLong, E. R., DeLong, D. M. and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44, 837–845.

Di Giuseppe, F., Molteni, F. and Tompkins, A. M. (2013). A rainfall calibration methodology for impacts modelling based on spatial mapping. *Quarterly Journal of the Royal Meteorological Society*, 139, 1389–1401.

Diebold, F. X., Gunther, T. A. and Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39, 863–883.

Dorfman, D. D. and Alf, E. (1969). Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals – rating-method data. *Journal of Mathematical Psychology*, 6, 487–496.

Ebert, E. E. (2001). Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Monthly Weather Review*, 129, 2461–2480.

Egan, J. P. (1975). *Signal Dectection Theory and ROC Analysis*. Academic Press, New York.

Egan, J. P., Greenberg, G. Z. and Schulman, A. I. (1961). Operating characteristics, signal detectability, and the method of free response. *The Journal of the Acoustical Society of America*, 33, 993–1007.

Ehm, W., Gneiting, T., Jordan, A. and Krüger, F. (2016). Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings. *Journal of the Royal Statistical Society Series B*, 78, 505–562.

Engel, T., Fink, A. H., Knippertz, P., Pante, G. and Bliefernicht, J. (2017). Extreme precipitation in the West African cities of Dakar and Ouagadougou: Atmospheric dynamics and implications for flood risk assessments. *Journal of Hydrometeorology*, 18, 2937–2957.

Etzioni, R., Pepe, M., Longton, G., Hu, C. and Goodman, G. (1999). Incorporating the time dimension in receiver operating characteristic curves: A case study of prostate cancer. *Medical Decision Making*, 19, 242–251.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27, 861–874.

Fawcett, T. and Niculescu-Mizil, A. (2007). PAV and the ROC convex hull. *Machine Learning*, 68, 97–106.

Fiebrich, C. A. and Crawford, K. C. (2001). The impact of unique meteorological phenomena detected by the Oklahoma Mesonet and ARS Micronet on automated quality control. *Bulletin of the American Meteorological Society*, 82, 2173–2187.

Fink, A. H., Engel, T., Ermert, V., van der Linden, R., Schneidewind, M., Redl, R., Afiesimama, E., Thiaw, W. M., Yorke, C., Evans, M. and Janicot, S. (2017). Mean climate and seasonal cycle. In *Meteorology of Tropical West Africa: The Forecasters' Handbook* (D. J. Parker and M. Diop-Kane, eds.), chap. 1. Wiley-Blackwell, Chichester, 1–39.

Fink, A. H., Parker, D. J., Lafore, J.-P., Ngamini, J.-B., Afiesimama, E., Beljaars, A., Bock, O., Christoph, M., Faccani, C., Karbou, F., Polcher, J., Mumba, Z., Nuret, M., Pohle, S., Rabier, F., Tompkins, A. M. and Wilson, G. (2011). Operational meteorology in West Africa: Observational networks, weather analysis and forecasting. *Atmospheric Science Letters*, 12, 135–141.

Fink, A. H. and Reiner, A. (2003). Spatiotemporal variability of the relation between African easterly waves and West African squall lines in 1998 and 1999. *Journal of Geophysical Research*, 108, 1–17.

Fink, A. H., Vincent, D. G. and Ermert, V. (2006). Rainfall types in the West African Sudanian zone during the summer monsoon 2002. *Monthly Weather Review*, 134, 2143–2164.

Fortin, V., Favre, A.-C. and Saïd, M. (2006). Probabilistic forecasting from ensemble prediction systems: Improving upon the best-member method by using a different weight and dressing kernel for each member. *Quarterly Journal of the Royal Meteorological Society*, 132, 1349–1369.

Fraley, C., Raftery, A. E. and Gneiting, T. (2010). Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Monthly Weather Review*, 138, 190–202.

Friederichs, P. and Thorarinsdottir, T. L. (2012). Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics*, 23, 579–594.

Geiger, R. (1961). *Überarbeitete Neuausgabe von Geiger, R. Köppen-Geiger / Klima der Erde. (Wandkarte 1:16 Mill.)*. Klett-Perthes, Gotha.

Gneiting, T. (2008). Editorial: Probabilistic forecasting. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 171, 319–321.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106, 746–762.

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B*, 69, 243–268.

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1, 125–151.

Gneiting, T. and Raftery, A. E. (2005). Weather forecasting with ensemble methods. *Science*, 310, 248–249.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102, 359–378.

Gneiting, T., Raftery, A. E., Westveld, A. H. and Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133, 1098–1118.

Gneiting, T. and Ranjan, R. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, 7, 1747–1782.

Gneiting, T. and Vogel, P. (2018). *Receiver Operating Characteristic (ROC) curves*. Preprint, `arXiv:1809.04808`.

Hagedorn, R., Buizza, R., Hamill, T. M., Leutbecher, M. and Palmer, T. N. (2012). Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 138, 1814–1827.

Haiden, T., Rodwell, M. J., Richardson, D. S., Okagaki, A., Robinson, T. and Hewson, T. (2012). Intercomparison of global model precipitation forecast skill in 2010/11 using the SEEPS score. *Monthly Weather Review*, 140, 2720–2733.

Hamill, T. M. and Colucci, S. J. (1997). Verification of Eta–RSM short-range ensemble forecasts. *Monthly Weather Review*, 125, 1312–1327.

Hamill, T. M. and Juras, J. (2006). Measuring forecast skill: Is it real skill or is it the varying climatology? *Quarterly Journal of the Royal Meteorological Society*, 132, 2905–2923.

Hamill, T. M., Snyder, C. and Morss, R. E. (2000). A comparison of probabilistic forecasts from bred, singular-vector, and perturbed observation ensembles. *Monthly Weather Review*, 128, 1835–1851.

Hanley, A. and McNeil, J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143, 29–36.

Hanley, J. A. and McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148, 839–843.

Heagerty, P. J., Lumley, T. and Pepe, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, 56, 337–344.

Hemri, S., Scheuerer, M., Pappenberger, F., Bogner, K. and Haiden, T. (2014). Trends in the predictive performance of raw ensemble weather forecasts. *Geophysical Research Letters*, 41, 9197–9205.

Hilden, J. (1991). The area under the ROC curve and its competitors. *Medical Decision Making*, 11, 95–101.

Hilden, J. and Gerds, T. A. (2014). A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. *Statistics in Medicine*, 33, 3405–3414.

Hirpa, F. A., Gebremichael, M. and Hopson, T. (2010). Evaluation of high-resolution satellite precipitation products over very complex terrain in Ethiopia. *Journal of Applied Meteorology and Climatology*, 49, 1044–1051.

Houze, R. A., Rasmussen, K. L., Zuluaga, M. D. and Brodzik, S. R. (2015). The variable nature of convection in the tropics and subtropics: A legacy of 16 years of the Tropical Rainfall Measuring Mission satellite. *Reviews of Geophysics*, 53, 994–1021.

Hsieh, F. and Turnbull, B. W. (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The Annals of Statistics*, 24, 25–40.

Huffman, G. J., Bolvin, D. T., Nelkin, E. J., Wolff, D. B., Adler, R. F., Gu, G., Hong, Y., Bowman, K. P. and Stocker, E. F. (2007). The TRMM multisatellite precipitation analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *Journal of Hydrometeorology*, 8, 38–55.

Jordan, A., Krueger, F. and Lerch, S. (2018). Evaluating probabilistic forecasts with scoringRules. *Journal of Statistical Software*. In press.

Kharin, V. V. and Zwiers, F. W. (2003). On the ROC score of probability forecasts. *Journal of Climate*, 16, 4145–4150.

Klar, M. (2017). *Statistical forecasts of rain occurrence over West Africa*. Master's thesis, Institute for Stochastics, Karlsruhe Institute of Technology.

Köppen, W. (1900). Versuch einer Klassifikation der Klimate, vorzugsweise nach ihren Beziehungen zur Pflanzenwelt. *Geographische Zeitschrift*, 6, 593–611.

Kottek, M., Grieser, J., Beck, C., Rudolf, B. and Rubel, F. (2006). World map of the Köppen-Geiger climate classification updated. *Meteorologische Zeitschrift*, 15, 259–263.

Krzanowski, W. J. and Hand, D. J. (2009). *ROC Curves for Continuous Data*. CRC Press, Boca Raton.

Krzysztofowicz, R. and Long, D. (1990). Fusion of detection probabilities and comparison of multisensor systems. *IEEE Transactions on Systems, Man, and Cybernetics*, 20, 665–677.

Lafore, J. P., Chapelon, N., Diop-Kane, M., Gueye, B., Largeron, Y., Lepape, S., Ndiaye, O., Parker, D. J., Poan, E., Roca, R., Roehrig, R. and Taylor, C. (2017). Deep convection. In *Meteorology of Tropical West Africa: The Forecasters' Handbook* (D. J. Parker and M. Diop-Kane, eds.), chap. 3. Wiley-Blackwell, Chichester, 90–129.

Leith, C. E. (1974). Theoretical skill of Monte Carlo forecasts. *Monthly Weather Review*, 102, 409–418.

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F. and Gneiting, T. (2017). Forecaster's dilemma: Extreme events and forecast evaluation. *Statistical Science*, 32, 106–127.

Leutbecher, M. and Palmer, T. N. (2008). Ensemble forecasting. *Journal of Computational Physics*, 227, 3515–3539.

Levasseur, K. M. (1984). A probabilistic proof of the Weierstrass approximation theorem. *The American Mathematical Monthly*, 91, 249–250.

Little, M. A., McSharry, P. E. and Taylor, J. W. (2009). Generalized linear models for site-specific density forecasting of UK daily rainfall. *Monthly Weather Review*, 137, 1029–1045.

Lloyd, C. J. (2002). Estimation of a convex ROC curve. *Statistics and Probability Letters*, 59, 99–111.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20, 130–141.

Maggioni, V., Meyers, P. C. and Robinson, M. D. (2016). A review of merged high-resolution satellite precipitation product accuracy during the Tropical Rainfall Measuring Mission (TRMM) era. *Journal of Hydrometeorology*, 17, 1101–1117.

Mammen, E. (1991). Estimating a smooth monotone regression function. *The Annals of Statistics*, 19, 724–740.

Maranan, M., Fink, A. H. and Knippertz, P. (2018). Rainfall types over southern West Africa: Objective identification, climatology and synoptic environment. *Quarterly Journal of the Royal Meteorological Society*, 144, 1628–1648.

Marsham, J. H., Dixon, N. S., Garcia-Carreras, L., Lister, G. M. S., Parker, D. J., Knippertz, P. and Birch, C. E. (2013). The role of moist convection in the West African monsoon system – insights from continental-scale convection-permitting simulations. *Geophysical Research Letters*, 40, 1843–1849.

Mason, S. J. and Graham, N. E. (2002). Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society*, 128, 2145–2166.

Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, 22, 1087–1096.

Mathon, V., Laurent, H. and Lebel, T. (2002). Mesoscale convective system rainfall in the Sahel. *Journal of Applied Meteorology*, 41, 1081–1092.

Metz, C. E., Herman, B. A. and Shen, J.-H. (1998). Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Statistics in Medicine*, 17, 1033–1053.

Metz, C. E. and Kronman, H. B. (1980). Statistical significance tests for binormal ROC curves. *Journal of Mathematical Psychology*, 22, 218–243.

Millar, P. W. (1984). A general approach to the optimality of minimum distance estimators. *Transactions of the American Mathematical Society*, 286, 377–418.

Miller, M., Buizza, R., Haseler, J., Hortal, M., Janssen, P. and Untch, A. (2010). Increased resolution in the ECMWF deterministic and ensemble prediction systems. *ECMWF newsletter*, 124, 10–16.

Molteni, F., Buizza, R., Palmer, T. N. and Petroliagis, T. (1996). The ECMWF ensemble prediction system: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, 122, 73–119.

Murphy, A. H. (1977). The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. *Monthly Weather Review*, 105, 803–816.

Murphy, A. H. and Winkler, R. L. (1977). Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 26, 41–47.

Murphy, A. H. and Winkler, R. L. (1987). A general framework for forecast verification. *Monthly Weather Review*, 115, 1330–1338.

Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer, New York.

Nesbitt, S. W., Cifelli, R. and Rutledge, S. A. (2006). Storm morphology and rainfall characteristics of TRMM precipitation features. *Monthly Weather Review*, 134, 2702–2721.

Palmer, T. N. (2000). Predicting uncertainty in forecasts of weather and climate. *Reports on Progress in Physics*, 63, 71–116.

Pantillon, F., Knippertz, P., Marsham, J. H. and Birch, C. E. (2015). A parameterization of convective dust storms for models with mass-flux convection schemes. *Journal of the Atmospheric Sciences*, 72, 2545–2561.

Park, Y. Y., Buizza, R. and Leutbecher, M. (2008). TIGGE: Preliminary results on comparing and combining ensembles. *Quarterly Journal of the Royal Meteorological Society*, 134, 2029–2050.

Parker, D. J., Fink, A. H., Janicot, S., Ngamini, J.-B., Douglas, M., Afiesimama, E., Agusti-Panareda, A., Beljaars, A., Dide, F., Diedhiou, A., Lebel, T., Polcher, J., Redelsperger, J.-L., Thorncroft, C. and Wilson, G. A. (2008). The AMMA radiosonde program and its implications for the future of atmospheric monitoring over Africa. *Bulletin of the American Meteorological Society*, 89, 1015–1027.

Pearson, K. J., Lister, G. M. S., Birch, C. E., Allan, R. P., Hogan, R. J. and Woolnough, S. J. (2014). Modelling the diurnal cycle of tropical convection across the 'grey zone'. *Quarterly Journal of the Royal Meteorological Society*, 140, 491–499.

Pepe, M. S. (2000). An interpretation for the ROC curve and inference using GLM procedures. *Biometrics*, 56, 352–359.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, Oxford.

Pesce, L. L., Metz, C. E. and Berbaum, K. S. (2010). On the convexity of ROC curves estimated from radiological test results. *Academic Radiology*, 17, 960–968.

Pinson, P. and Hagedorn, R. (2012). Verification of the ECMWF ensemble forecasts of wind speed against analyses and observations. *Meteorological Applications*, 19, 484–500.

Pohl, B., Janicot, S., Fontaine, B. and Marteau, R. (2009). Implication of the Madden–Julian oscillation in the 40-day variability of the West African monsoon. *Journal of Climate*, 22, 3769–3785.

R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Raftery, A. E., Gneiting, T., Balabdaoui, F. and Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133, 1155–1174.

Richard, E., Buzzi, A. and Zängl, G. (2007). Quantitative precipitation forecasting in the Alps: The advances achieved by the Mesoscale Alpine Programme. *Quarterly Journal of the Royal Meteorological Society*, 133, 831–846.

Richardson, L. F. (1922). *Weather Prediction by Numerical Methods*. Cambridge University Press, Cambridge.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C. and Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12, 1–8.

Roca, R., Chambon, P., Jobard, I., Kirstetter, P. E., Gosset, M. and Bergés, J. C. (2010). Comparing satellite and surface rainfall products over West Africa at meteorologically relevant scales during the AMMA campaign using error estimates. *Journal of Applied Meteorology and Climatology*, 49, 715–731.

Rodwell, M. J., Richardson, D. S., Hewson, T. D. and Haiden, T. (2010). A new equitable score suitable for verifying precipitation in numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 136, 1344–1363.

Rüschendorf, L. (2009). On the distributional transform, Sklar's theorem, and the empirical copula process. *Journal of Statistical Planning and Inference*, 139, 3921–3927.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66, 783–801.

Scheuerer, M. (2014). Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society*, 140, 1086–1096.

Scheuerer, M. and Hamill, T. M. (2015). Statistical post-processing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Monthly Weather Review*, 143, 4578–4596.

Shorack, G. R. and Wellner, J. A. (2009). *Empirical Processes with Applications to Statistics*. SIAM, Philadelphia.

Sing, T., Sander, O., Beerenwinkel, N. and Lengauer, T. (2005). ROCR: Visualizing classifier performance in R. *Bioinformatics*, 21, 3940–3941.

Sloughter, J. M., Raftery, A. E., Gneiting, T. and Fraley, C. (2007). Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly Weather Review*, 135, 3209–3220.

Söhne, N., Chaboureau, J.-P. and Guichard, F. (2008). Verification of cloud cover forecast with satellite observation over West Africa. *Monthly Weather Review*, 136, 4421–4434.

Stephens, G. L., L'Ecuyer, T., Forbes, R., Gettelmen, A., Golaz, J.-C., Bodas-Salcedo, A., Suzuki, K., Gabriel, P. and Haynes, J. (2010). Dreary state of precipitation in global models. *Journal of Geophysical Research: Atmospheres*, 115, D24211.

Swets, J. A. (1973). The relative operating characteristic in psychology. *Science*, 182, 990–1000.

Swets, J. A. (1986). Indices of discrimination or diagnostic accuracy. *Psychological Bulletin*, 99, 100–117.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, 240, 1285–1293.

Swinbank, R., Kyouda, M., Buchanan, P., Froude, L., Hamill, T. M., Hewson, T. D., Keller, J. H., Matsueda, M., Methven, J., Pappenberger, F., Scheuerer, M., Titley, H. A., Wilson, L. and Yamaguchi, M. (2016). The TIGGE project and its achievements. *Bulletin of the American Meteorological Society*, 97, 49–67.

Talagrand, O., Vautard, R. and Strauss, B. (1997). Evaluation of probabilistic prediction systems, paper presented at ECMWF Workshop on Predictability, Eur. Cent. for Med. Range Weather Forecasts. Reading, UK.

Thorarinsdottir, T. L. and Gneiting, T. (2010). Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression. *Journal of the Royal Statistical Society Series A*, 173, 371–388.

Toth, Z. and Kalnay, E. (1993). Ensemble forecasting at NMC: The generation of perturbations. *Bulletin of the American Meteorological Society*, 74, 2317–2330.

van der Linden, R., Fink, A. H., Pinto, J. G. and Phan-Van, T. (2017). The dynamics of an extreme precipitation event in northeastern Vietnam in 2015 and its predictability in the ECMWF ensemble prediction system. *Weather and Forecasting*, 32, 1041–1056.

Venkatraman, E. S. (2000). A permutative test to compare receiver operating characteristic curves. *Biometrics*, 56, 1134–1138.

Venkatraman, E. S. and Begg, C. B. (1996). A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. *Biometrika*, 83, 835–848.

Vizy, E. K. and Cook, K. H. (2014). Impact of cold air surges on rainfall variability in the Sahel and wet African tropics: A multi-scale analysis. *Climate Dynamics*, 43, 1057–1081.

Vogel, P., Knippertz, P., Fink, A. H., Schlueter, A. and Gneiting, T. (2018). Skill of global raw and postprocessed ensemble predictions of rainfall over northern tropical Africa. *Weather and Forecasting*, 33, 369–388.

Webster, P., Toma, V. E. and Kim, H.-M. (2011). Were the 2010 Pakistan floods predictable? *Geophysical Research Letters*, 38, L04806.

Webster, P. J. (2013). Meteorology: Improve weather forecasts for the developing world. *Nature*, 493, 17.

Wheeler, M. and Kiladis, G. N. (1999). Convectively coupled equatorial waves: Analysis of clouds and temperature in the wavenumber–frequency domain. *Journal of the Atmospheric Sciences*, 56, 374–399.

Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences*. 3rd ed. Academic Press, Amsterdam.

Young, M. P., Chiua, J. C., Williams, C. J. R., Stein, T. H. M., Stengel, M., Fielding, M. D. and Black, E. (2018). Climatology and diurnal variability of warm rain events over southern West Africa from geostationary satellite observations for climate monitoring and model evaluation. *Quarterly Journal of the Royal Meteorological Society*. In press.

Zhou, X.-H., Obuchowski, N. A. and McClish, D. K. (2011). *Statistical Methods in Diagnostic Medicine*. 2nd ed. Wiley, Hoboken.

Zou, K. H. and Hall, W. J. (2000). Two transformation models for estimating an ROC curve derived from continuous data. *Journal of Applied Statistics*, 27, 621–631.

Zou, K. H., Resnic, F. S., Talos, I. F., Goldberg-Zimring, D., Bhagwat, J. G., Haker, S. J., Kikinis, R., Jolesz, F. A. and Ohno-Machado, L. (2005). A global goodness-of-fit test for receiver operating characteristic curve analysis via the bootstrap method. *Journal of Biomedical Informatics*, 38, 395–403.

Zweig, M. H. and Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39, 561–577.