Data-Driven Agroclimate Modeling and Forecasting Based on Earth Observations and Predictions: A Study of Evapotranspiration, Precipitation, and Crop Yields

by

Hanoi Medina González

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
May 2, 2020

Keywords: numerical weather predictions, remote sensing, MODIS, forecast post-processing, machine learning

Approved by

Dr. Di Tian, Chair, Assistant Professor, Dep. of Crop, Soil, and Environ. Sciences
Dr. Brenda Ortiz, Professor & Extension Specialist, Dep. of Crop, Soil and Environ. Sciences
Dr. Puneet Srivastava, Associate Dean for Research and Associate Director of Maryland Agricultural Experiment Station (MAES), University of Maryland
Dr. Willian Batchelor, Professor, Dep. of Biosystems Engineering

ABSTRACT

Meeting the growing food demand in a sustainable manner is the main challenge faced by agriculture, at a time when the climate changes increasingly threaten food security. Innovative agro-climate approaches are needed to convert the growing amount of information emerging thanks to technological advances into products that help to adopt better decisions. Since water intervenes much of the climate change impacts on agriculture, accurately forecasting of precipitation and evapotranspiration is of the uppermost importance for minimizing the effect of adverse weather. The improvements of the numerical weather prediction (NWP) models and the statistical post-processing techniques provide unprecedented opportunities to better anticipate the changes in precipitation and evapotranspiration. The use of satellite remote sensing techniques for in-season forecasting of major crop yields over large areas is also of special interest, as it provides proxies of food security and food prices. While datasets from the moderate resolution imaging spectroradiometer (MODIS) are advantageous for crop forecasting, more research is needed on how factors such as the type of MODIS product, the statistical model or the training domain affect the crop yield forecasts. This study has been aimed to develop and evaluate new data-driven approaches for agro-climate forecasting, which combines NWP forecasts, remote sensing data, numerical modeling and machine learning techniques for improving crop water demand and crop yields forecasting in agricultural ecosystems. The manuscript is divided into six main chapters. In Chapter I, I provide a general introduction of the research. In Chapter II, single and multimodel ensemble forecasts of daily reference evapotranspiration, based on three leading NWP models over the continental U.S (CONUS), are produce and evaluated. The ability of three states of art probabilistic post-processing methods for improving NWP-based daily and weekly reference evapotranspiration forecasts over the CONUS in evaluated in Chapter III. In Chapter IV I evaluate leading NWP models and post-processing methods for improving ensemble precipitation forecasts over Brazil. In Chapter V it is evaluated a new optimization framework for the MODIS-based county and state-level corn yield forecasting over major producing states of the U.S. Finally, Chapter VI provides concluding remarks. The results represent a step forward in the efforts for improving precipitation, evapotranspiration and crop yield forecasting across multiple scales.

ACKNOWLEDGMENTS

I would like to express my sincere appreciation to my advisor, Dr. Di Tian, for the guidance, support and confidence throughout this PhD degree project. Without his leadership the outcomes would not be what they are. Thanks are also extended to the members of my committee: Dr. Brenda Ortiz, Dr. Puneet Srivastava, and Dr. Willian Batchelor for their suggestions and support whenever I needed it.

I want to thank this country, for hosting me and my family, and for the opportunities for pursuing my goals. My gratitude to Auburn University staff, and in particular to the Department of Crop, Soil and Environmental Sciences staff for their generous support over these years. Thanks also to the colleagues and friends I made in Auburn during this chapter of my life.

Very special thanks to my loved wife Yamilet and son Alvaro. Our recent history is that of the three (immigrant) Musketeers: all for one and one for all. Thanks to them this journey has been easier and gratifying. Special thanks also to my mother Eloisa and my father Juan, for their infinite love and for enduring the sacrifice I imposed them.

TABLE OF CONTENTS

LIST OF FIGURES

replacing one observed variable, either temperature (*T*), relative humidity (*RH*), solar radiation (*Rs*), or wind speed (*u*) at a time with one corresponding forecast.

Figure III.1. Relative mean error (rME), relative root mean square error (rRMSE), and correlation considering daily forecasts for different lead times over the SE and NW regions.

Figure III.2. Binned spread-skill plots accounting for the mean of the ensemble standard deviation deciles against the mean RMSE of the forecasts in each decile over the verification period based on all pairs of forecasts and observations at a) 1-day and b) 7-day lead. The panel in the right and the bottom shows the corresponding rank histograms. The correlation between the standard deviations and the absolute errors is reported after the colon. The solid line represents the 1:1 relationship.

Figure III.3. a) BSS for every region and lead time of the daily ECMWF forecasts post-processed using simple bias correction (used as reference BSS values) and b-e) differences between the BSS of the daily ECMWF forecasts post-processed with the b) NGR and c) AKD methods and the daily ECMWF-NCEP-UKMO forecasts post-processed with the d) NGR and e) BMA methods and the reference BSS.

Figure III.4. Regional mean weight coefficient b of the NGR technique (left panel) and the weight coefficient w of the BMA technique (right panel) for the post-processed daily ECMWF-NCEP-UKMO forecasts at different lead days.

Figure III.5. Whisker plot with the 2.5^{th}, 25^{th}, 50^{th}, 75^{th} and 97.5^{th} percentile of the distribution of the rME, rRMSE and correlation of weekly forecasts across different regions.

Figure III.6. Binned spread-skill plots for the weekly forecasts accounting for the mean of the ensemble standard deviation deciles against the mean RMSE of the forecasts in each decile over the verification period using all pairs of forecasts and observations. The panel in the right and the bottom shows the corresponding rank histograms. The correlation between the standard deviations and the absolute errors is included in the legend. The solid line represents the 1:1 relationship.

Figure III.7. Comparison between BC and NGR based Brier Skill Scores considering a) ECMWF and ECMWF-UKMO forecasts, b) NCEP, and c) UKMO forecasts across the different climate regions.

Figure IV.1. Regions of Brazil involved in this study corresponding to the six major natural biomes as defined in IBGE (2016).

Figure IV.2. Cumulative distribution of the precipitations higher than 1 mm over 1985-2010.

Figure IV.3. a) correlation and RMSE (mm) of the NCEP precipitation raw forecasts at each biome in January (1), April (4), July (7) and October (10), for lead times 1.5, 3.5 and 5.5; and b) differences between correlation and RMSE of the ECMWF raw forecasts as well as the Control analog forecasts and the NCEP raw forecasts.

Figure V.6. Distribution of the annual mean absolute percentage error (MAPE) and annual R2 of the county-level forecasts considering the different spatial domains with the linear regression based on $LAI_{EVI}\langle\overline{209}\rangle$ and the elastic net and the random forest based on $LAI_{EVI}\langle\overline{193}, \overline{201}, \overline{209}\rangle$. The bars in the top of the plots denote significant differences with respect to the domain aligned with the mark on the left, resulting from a pairwise Wilcoxon test analysis with Bonferroni adjustments.

Figure V.7. Distribution of percent error (PE) and R^2 of the annual **state-level** forecasts considering the linear regression based on $LAI_{EVI}\langle\overline{209}\rangle$ and the elastic net and random forests based on $LAI_{EVI}\langle\overline{193}, \overline{201}, \overline{209}\rangle$ for the four spatial domains as well as the percent considering the NASS forecasts. The bars in the top of the plots denote significant differences with respect to the domain aligned with the mark on the left.

Figure V.8. Median of the differences between the MAPE, the R^2 (at county level) and the PE (at state-level) with elastic net and random forest and the MAPE, the R^2 and the PE with linear regression for the four model domains. The subsets of LAI_{EVI} used by elastic net and random forest are indicated at the top. The linear regressions based on $\langle\overline{209}\rangle$ and the best domains identified in Table V.2.

Figure V.9. The top panel shows the cumulative probability distribution of the differences between the percent errors (PE) at the county-level considering the LAI_{NDVI}, LAI and FPAR based forecasts and the PE considering the LAI_{EVI} based forecasts. The bottom panel shows the distribution of the differences in the annual R^2. The comparisons considered the raw composite's temporal resolution and the four spatial domains (except the global domain in Nebraska, which provided exceptionally poor performance).

Figure V.10. Boxplots in the upper panel show the differences in the percent errors (PE) using the LAI_{NDVI}, LAI and FAPAR $\langle\overline{209}\rangle$ with respect to PE using LAI_{EVI} $\langle\overline{209}\rangle$ and the PE using the NASS forecasts. Bar plots in the bottom show the corresponding differences in R^2.

Figure V.11. Error map based on linear regression on $\langle\overline{209}\rangle$ considering the best MODIS products (NDVI for Iowa and EVI for the rest and the best domains at county-level (based on Table V.2: **Global** in Illinois, Indiana, and Iowa, **County** in Nebraska and **District** in Ohio). Polygons in dark gray indicate missing values. The map indicated with an arrow shows the county with the lowest mean percent error across years in each state.

Figure V.A1. Distribution of annual mean absolute percentage error (MAPE) and annual R^2 of the county-level forecasts for the different spatial domains using the linear regression based on $\langle 209\rangle$. The bars in the top of the plots denote significant differences with respect to the domain aligned with the mark on the left.

Figure V.A2. Distribution of annual mean absolute percentage error (MAPE) and annual R^2 of the **county-level** forecasts for the different spatial domains using the elastic net based on a) $\langle\overline{161},\overline{177},\overline{193},\overline{209}\rangle$, b) $\langle\overline{153}\text{ to }\overline{209}, \text{in steps of } 8\rangle$, c) $\langle 193, 201, 201\rangle$ and d) $\langle 153 \text{ to } 209, \text{in steps of } 8\rangle$. The bars in the top of the plots denote significant differences with respect to the domain aligned with the mark on the left.

Figure V.A3. Distribution of annual mean absolute percentage error (MAPE) and annual R^2 of the county-level forecasts for the different spatial domains using the random forest based on a) $\langle\overline{161},\overline{177},\overline{193},\overline{209}\rangle$, b) $\langle\overline{153}\text{ to }\overline{209}, \text{in steps of } 8\rangle$, c) $\langle 193, 201, 201\rangle$ and d) $\langle 153 \text{ to } 209, \text{in steps of } 8\rangle$. The bars in the top of the plots denote significant differences with respect to the domain aligned with the mark on the left.

Figure V.B1. Distribution of percent error (PE) and R^2 of the annual **state-level** forecasts for the different spatial domains using the linear regression based on $LAI_{EVI}\langle 209\rangle$ and the PE and R^2 with the NASS forecasts. The bars in the top of the plots denote significant differences with respect to the domain aligned with the mark on the left.

Figure V.B2. Distribution of percent error (PE) and R^2 of the annual **state-level** forecasts for the different spatial domains using the elastic net with the LAI_{EVI} based on a) $\langle\overline{161},\overline{177},\overline{193},\overline{209}\rangle$, b) $\langle\overline{153}\text{ to }\overline{209}, \text{in steps of } 8\rangle$, c) $\langle 193, 201, 209\rangle$ and d) $\langle 153 \text{ to } 209, \text{in steps of } 8\rangle$, as well as the PE and R^2 with the NASS forecasts. The bars in the top of the plots denote significant differences with respect to the domain aligned with the mark on the left.

Figure V.B3. Distribution of percent error (PE) and R^2 of the annual **state-level** forecasts for the different spatial domains using the random forest with the LAI_{EVI} based on a) $\langle\overline{161},\overline{177},\overline{193},\overline{209}\rangle$, b) $\langle\overline{153}\text{ to }\overline{209}, \text{in steps of } 8\rangle$, c) $\langle 193, 201, 209\rangle$ and d) $\langle 153 \text{ to } 209, \text{in steps of } 8\rangle$ as well as the PE and R^2 with the NASS forecasts.

Figure V.C1. Idem to Fig. V.7, but for elastic net and random forest regressions based on subsets of the LAI_{EVI} (indicated on the top of the figure) at high temporal resolution.

LIST OF TABLES

Table V.3. Mean differences between the county-level MAPE and R^2 based on the composites at daily resolution and the MAPE and R^2 based on the coarse temporal resolution for the three techniques (linear regression (LR), elastic net (EN) and random forest (RF)), and the four spatial domains (global, state, district, and county-based). The "*", "**" and "***" indicates that the differences are significant at α confidence levels 0.95, 0.99 and 0.999 based on a Wilcoxon test.

LIST OF ABBREVIATIONS

100_Ens	A variant of the analog method
A_05pr_05pw	A variant of the analog method
A_09pr_01pw i	A variant of the analog method
ACCESS-G	Australian Community Climate and Earth-System Simulator
AKD	Affine Kernel Dressing
BC	Bias correction
BMA	Bayesian Model Averaging
BS	Brier Score
BSS	Brier Skill Score
C	Central
C. Analog	States for Control Analog (variant to the analog method)
CONUS	Continental U.S.
CRPS	Continuos Rank Probability Score
DOY	Day of the year
ECMWF	European Centre for Medium-Range Weather Forecasts
EN	Elastic Net
ENC	East North Central
ET_0	Reference crop evapotranspiration
EVI	Enhanced vegetation index
FAO	Food and Agriculture Organization
FPAR	Fraction of absorbed photosynthetic active radiation
GEFS	Global Ensemble Forecast System
GFS	Global Forecast System
LAI	Leaf Area Index
LNR	Linear Regression
LogF	A variant of the analog method
LR	Logistic Regression
MAPE	mean absolute percentage error
ME	Mean error
MODIS	Moderate-Resolution Imaging Spectroradiometer
MOS	Model Output Statistics
NASS	National Agricultural Statistics Service
NCEI	National Centers for Environmental Information
NCEP	National Centers for Environmental Prediction
NDVI	Normalized difference vegetation index
NE	North East
NGR	Non Homogeneous Gaussian Regression
NIR	Near Infrared

NW	North West
NWP	Numerical Weather Predictions
PDF	Probability density function
PE	Percent error
PM	Penman-Monteith
R2	Determination Coefficient
RF	Random Forest
RH	Relative Humidity
rME	Relative ME
RMSE	Root mean square error
rRMSE	Relative RMSE
Rs	Solar Radiation
S	South
SE	South East
Short_reg	A variant of the analog method
SW	South West
T	Temperature
THORPEX	The Observing System Research and Predictability Experiment
TIGGE	THORPEX Interactive Grand Global Ensemble
u	Wind speed
U.S.	United States
UKMO	United Kingdom Meteorological Office
USCRN	United States Climate Reference Network
UTC	Coordinated Universal Time
W	West
WNC	West North Central

CHAPTER I: GENERAL INTRODUCTION

1 BACKGROUND

Meeting the growing food demand sustainably is the main challenge faced by agriculture (FAO, 2017), at a time when the world population is expected to reach almost 10 billion by 2050 (United Nations, 2017), and the pressure on already limited natural resources and the adverse effects and uncertainty of climate changes deepen. The ongoing climate changes, in particular, represent an increasing threat for food security in the coming years because of the potential negative impacts on most agriculture ecosystems, including crop monocultures, grazing systems, arid-land pastoral systems, etc. The projected increases in temperature and the changes in the precipitation regimes, including the increments in the frequency and intensity of the extreme events such as droughts and floods (Collins et al., 2013) jeopardize crop (e.g. Takle and Gutowski, 2020), livestock (Rojas-Downing et al., 2018) and fish stocks (Brander, 2010) productivity. Hence, there is a growing demand for innovative systems that help to improve resource management and the decision-making, as a way to improve productivity and sustainability of agricultural ecosystems.

The advances in the information and computer technologies, including the improvements in internet connectivity, high-performance computing, satellite capabilities, and others; the huge increments in data availability both from ground and remote sensing sources; as well the progress in data analytics and open-source software packages, provide unparalleled opportunities for enhancing the efficiency in agricultural production (Janssen et al., 2017). In particular, the considerable improvements in the numerical weather prediction (NWP) models (Bauer et al., 2015) can be useful for mitigating the impacts and minimize losses because of the adverse weather, while can provide financial profits in sectors such agriculture. Since most aspects of crop culture are impacted by weather, the use of skilled weather forecasts may be of great help in agriculture (Sivakumar, 2006). The advances in the satellite and drone remote sensing capabilities have also largely enhanced our ability to assist management, monitoring and controlling activities in agriculture since they for example provide real-time data regarding in-season crop growth and development (e.g. Atzberger et al., 2013). However, the greater accessibility to information and technologies itself does not automatically generate products that are comprehensible and

appropriate to decision-making, but rather intelligent processing and analytics are needed for so. In particular, new agro-climate tools are needed to aid actors such as farmers, stakeholders, and rangers to cope with climate impacts.

Studies considering agroclimate models show that uncertainty in climate explains a considerable portion of the total uncertainty (e.g. Challinor et al., 2009). The accuracy of weather-related forecasts is crucial for the success of any effort aimed to anticipate the impacts of climate variability and change on many agricultural activities. Because water mediates much of the climate change effects on agriculture, accurately forecasting of precipitation and soil and crop evapotranspiration is of the uppermost importance. The largest crop failures events in history have been commonly associated with anomalies in the balance between these water budget terms. Precipitation and evapotranspiration are two fundamental forcings of water status in agro-hydrological models, with evapotranspiration being often retrieved as a function of the daily reference crop evapotranspiration (ET_0), which is evapotranspiration from a well-watered hypothetical reference crop. However, the accurate representation of both precipitation and ET_0 in the models is typically challenging. Precipitation, in particular, is often considered as a highly uncertain model input (e.g., Renard et al., 2010), because it has short spatial and temporal correlation length scales.

Medium range weather forecasts of precipitation and reference evapotranspiration, enabling farmers to coordinate and implement suitable cultural operations, may be particularly useful for improving water management in agriculture. Medium-range forecasts cover a validity period that is more appropriate for addressing decisions than that considered by short-range forecasts, and at the same time they are more accurate than seasonal forecasts (Thielen et al., 2009). While efforts have been made to incorporate precipitation and reference evapotranspiration medium-range forecasts into management and planning of water resources, more research is needed for reducing the uncertainty of the assessments. For example, while several global NWP models are potentially useful for the forecast delivery, few studies have comprehensively compared the performance of different models at regional scales. The potential benefits derived from the use of forecast ensembles, in particular the use of multi-model ensembles of reference evapotranspiration, have been insufficiently explored in literature. The use of ensembles has been shown critical for envisaging the impacts of climate changes on crops and evaluating the uncertainty associated with crop models (Challinor et al., 2013). Finally, while several statistical

post-processing methods have shown effective for improving the performance of the NWP forecasts, more research is needed to evaluate the potential of state of art probabilistic methods, consenting adjustments in both the mean and the deviation of the model ensembles.

The remote sensing datasets are also useful to improve resource management and decision-making in the agricultural sector by capturing the multispectral dynamics of the land surface over across multiple scales. It has been proven powerful for biomass and yield estimation and forecasting (Becker-Reshef et al., 2010), water resource management (e.g., Thenkabai et al., 2009), land cover and crop type classification (Kussul et al., 2017), precision agriculture (Mulla, D.J., 2013), weeding control (Lamb et al. 2001), etc. The use of the satellite remote sensing technologies for the in-season crop yield forecasting across scales is of special interest since it provides proxies of food security and food prices. In particular, the crop yield forecasting of major cereal crops like corn and over large producer regions, as some states of the Midwest of the U.S., may have important implications for food security worldwide. Datasets based on the moderate resolution imaging spectroradiometer (MODIS) have been shown advantageous for crop yield forecasting across scales (e.g., Becker-Reshef et al., 2010; Mkhabela et al., 2011). MODIS instruments have monitored the land surface dynamics for about two decades with daily revisit frequency, providing an inestimable source of information for testing and validating the robustness of the crop yield forecasting frameworks. However, studies (e.g. Zhou and Zhang et al., 2016) that more research on how to optimally using MODIS datasets is needed. Comprehensively analyzes considering how factors such as the type of MODIS product, the regression model and the configuration (e.g. the length) of the time series of MODIS data affect the crop yield forecasting may be helpful but are lacking in the literature.

The PhD research has been aimed to develop and evaluate new data-driven approaches for agro-climate forecasting, which combines NWP forecasts, remote sensing data, numerical modeling and machine learning techniques for improving crop water demand and crop yields forecasting in agricultural ecosystems. The manuscript is divided into four main chapters (Chapters II-V) in addition to an introductory chapter (Chapter I) and a conclusion chapter (Chapter VI). The specific objectives of Chapters (II-V) are included in the following section.

2 OBJECTIVES

The specific objectives of this research are shown in Table I.1.

Table I.1. Specific objectives in chapters II-V.

Chapter	Objective(s)
II	• Produce and evaluate deterministic and probabilistic ET_0 forecasts from both single and multi-model combinations of ECMWF, NCEP and UKMO forecasts from the TIGGE dataset. • Evaluate the effects of different ensemble sizes on ET_0 forecast performance, as well as the impact of the individual weather forecast variable on ET_0 forecasts.
III	• Evaluate and compare multiple strategies for post-processing both daily and weekly ET_0 forecasts using the Non Homogeneous Gaussian Regression, Affine Kernel Dressing and Bayesian Model Averaging approaches.
IV	• Document the performance of the NCEP and ECMWF daily precipitation ensemble forecasts using Brazil as a study case. • Evaluate the NCEP-based precipitation forecasts post-processed using analog methods with different strategies. • Compare the performance of Analog-based methods with the Logistic Regression method.
V	• Construct an optimized framework for MODIS-based corn yield forecasts over major producer states of the U.S., by considering multiple machine learning techniques, product subsets, model domains, and temporal resolutions. • Evaluate and compare the performance of the optimized framework based on the MODIS NDVI, LAI, and FPAR products.

REFERENCES

1. Atzberger, C., 2013. Advances in remote sensing of agriculture: Context description, existing operational monitoring systems and major information needs. Remote sensing, 5(2), pp.949-981.
2. Brander, K., 2010. Impacts of climate change on fisheries. Journal of Marine Systems, 79(3-4), pp.389-402.
3. Challinor, A.J., Smith, M.S., Thornton, P., 2013. Use of agro-climate ensembles for quantifying uncertainty and informing adaptation. Agricultural and Forest Meteorology, 170, pp. 2-7.
4. Challinor, A.J., Wheeler, T., Hemming, D., Upadhyaya, H.D., 2009. Ensemble yield simulations: crop and climate uncertainties, sensitivity to temperature and genotypic adaptation to climate change. Climate Research, 38(2), pp.117-127.
5. FAO. 2017. The future of food and agriculture – Trends and challenges. Rome. Available at http://www.fao.org/3/a-i6583e.pdf, last access: 03/20/2020.
6. Collins, M., R. Knutti, J. Arblaster, J.-L. Dufresne, T. Fichefet, P. Friedlingstein, X. Gao, W.J. Gutowski, T. Johns, G. Krinner, M. Shongwe, C. Tebaldi, A.J. Weaver, M. Wehner, 2013: Long-term Climate Change: Projections, Commitments and Irreversibility. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.
7. Janssen, S.J., Porter, C.H., Moore, A.D., Athanasiadis, II.N., Foster, II., Jones, J.W., Antle, J.M., 2017. Towards a new generation of agricultural system data, models and knowledge products: Information and communication technology. Agricultural systems, 155, pp.200-212.
8. Kussul, N., Lavreniuk, M., Skakun, S., Shelestov, A., 2017. Deep learning classification of land cover and crop types using remote sensing data. IEEE Geoscience and Remote Sensing Letters, 14(5), pp.778-782.
9. Lamb, D.W., Brown, R.B., 2001. Pa—precision agriculture: Remote-sensing and mapping of weeds in crops. Journal of Agricultural Engineering Research, 78(2), pp.117-125.

10. Mulla, D.J., 2013. Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. Biosystems engineering, 114(4), pp.358-371.
11. Renard, B., Kavetski, D., Kuczera, G., Thyer, M., Franks, S.W., 2010. Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. Water Resources Research, 46(5).
12. Rojas-Downing, M.M., Nejadhashemi, A.P., Harrigan, T., Woznicki, S.A., 2017. Climate change and livestock: Impacts, adaptation, and mitigation. Climate Risk Management, 16, pp.145-163.
13. Sivakumar, M.V., 2006. Climate prediction and agriculture: current status and future challenges. Climate Research, 33(1), pp.3-17.
14. Takle, E.S., Gutowski Jr, W.J., 2020. Iowa's agriculture is losing its Goldilocks climate. PhT, 73(2), pp.26-33.
15. Thenkabail, P.S., Biradar, C.M., Noojipady, P., Dheeravath, V., Li, Y., Velpuri, M., Gumma, M., Gangalakunta, O.R.P., Turral, H., Cai, X. and Vithanage, J., 2009. Global irrigated area map (GIAM), derived from remote sensing, for the end of the last millennium. Int. J. Remote Sens, 30, 3679–3733.
16. United Nations, Department of Economic and Social Affairs, Population Division, 2017. World Population Prospects 2017 – Data Booklet (ST/ESA/SER.A/401). Available at https://population.un.org/wpp/Publications/Files/WPP2017_DataBooklet.pdf, last access: 03/20/2020.
17. Zhou, F., Zhang, A., 2016. Optimal subset selection of time-series MODIS images and sample data transfer with random forests for supervised classification modelling. Sensors, 16(11), p.1783.

CHAPTER II: MEDIUM-RANGE REFERENCE EVAPOTRANSPIRATION FORECASTS FOR THE CONTIGUOUS UNITED STATES BASED ON MULTI-MODEL NUMERICAL WEATHER PREDICTIONS

This chapter has been published in: *Journal of Hydrology*, 562, pp.502-517, 2018.

Abstract: Reference evapotranspiration (ET_0) plays a fundamental role in agronomic, forestry, and water resources management. Estimating and forecasting ET_0 have long been recognized as a major challenge for researchers and practitioners in these communities. This work explored the potential of multiple leading numerical weather predictions (NWPs) for estimating and forecasting summer ET_0 at 101 U.S. Regional Climate Reference Network stations over nine climate regions across the contiguous United States (CONUS). Three leading global NWP model forecasts from THORPEX Interactive Grand Global Ensemble (TIGGE) dataset were used in this study, including the single model ensemble forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF), the National Centers for Environmental Prediction Global Forecast System (NCEP), and the United Kingdom Meteorological Office forecasts (UKMO), as well as multi-model ensemble forecasts from the combinations of these NWP models. A regression calibration was employed to bias correct the ET_0 forecasts. Impact of individual forecast variables on ET_0 forecasts were also evaluated. The results showed that the ECMWF forecasts provided the least error and highest skill and reliability, followed by the UKMO and NCEP forecasts. The multi-model ensembles constructed from the combination of ECMWF and UKMO forecasts provided slightly better performance than the single model ECMWF forecasts. The regression process greatly improved ET_0 forecast performances, particularly for the regions involving stations near the coast, or with complex orography. The performance of ECMWF forecasts was only slightly influenced by the size of the ensemble members, particularly at short lead times. Even with fewer ensemble members, ECMWF still performed better than the other two NWPs. Errors in the radiation forecasts, followed by those in the wind, had the most detrimental effects on the ET_0 forecast performances.

1 INTRODUCTION

Accurate evapotranspiration (ET) forecasting is important for assessing agricultural water demand, driving hydrologic and crop simulation models, and enhancing agricultural and water management decision making. Evapotranspiration is often derived as a function of the daily reference crop evapotranspiration (ET_0), which is the evapotranspiration from a well-watered reference crop.

An internationally recognized standard method for computing ET_0 is the Penman-Monteith equation as specified by the Food and Agriculture Organization in the Irrigation and Drainage paper 56 (Allen et al., 1998). This method (hereinafter referred to as FAO-56 PM) is considered as one of the best methods for estimating daily ET_0 under different climate conditions. Since the FAO-56 PM equation is a physically-based approach, incorporating both physiological and aerodynamic parameters, it does not require any local calibration (e.g. Garcia et al., 2004). However, the FAO-56 PM equation requires the availability of a complete set of meteorological data including air temperature, wind speed, solar radiation and relative humidity.

Forecast outputs from numerical weather prediction (NWP) models can be used for ET_0 forecasting. The improvements in resolution, parameterization, and physical representation of the main processes and phenomena, has prompted the use of medium-range (1-10 days) NWP forecasts in many weather-dependent activities (Hamill et al., 2013). Forecast skill in the range from 3 to 10 days has been increasing by about one day per decade, meaning that today's 6-day forecast is as accurate as the 5-day forecast ten years ago (Bauer et al., 2015).

Medium-range forecasts are crucial for agronomy, forestry, and water resources management as they provide more time for decision making and planning compared with short-range forecasts, as well as producing considerably more accurate estimations than seasonal forecasts (Thielen et al., 2009). Nevertheless, implementing medium-range NWP forecasts is not straightforward. Correction methods are often needed to reduce forecast errors and account for local meteorological conditions that are not resolved at the spatial scale of the NWP model grid (e.g. Delle Monache et al., 2011; Glahn and Lowry, 1972; Gneiting, 2014; Gneiting et al., 2005; Pelosi et al., 2017; Wilks, 2006). Using multiple model ensembles, instead of a single model, and statistical post-processing of NWP models outputs are two of the several correction methods used to improve weather-related forecasts (Hamill, 2012). Hagedorn et al. (2012) found that the post-

processing procedure based on a simple bias correction approach can be particularly useful at locations affected by systematic errors, including areas with complex landscape or coastal grid points. Also, studies showed that the multi-model ensemble approach by combining multiple NWPs often have higher skill than any individual model due to compensation effects through combining models with different physics, numeric, and initial conditions (e.g. Hagedorn et al. 2005).

Recent studies have explored global numerical weather model outputs for forecasting medium-range ET_0 in real-time. For example, Perera et al. (2014) used ACCESS-G global model outputs to estimate ET_0 with lead times up to 9 days. This model is operated by the Australian Bureau of Meteorology with a spatial resolution of 80 km. In this study, the ET_0 forecasts showed an average RMSE less than 1 mm day^{-1} for lead time up to 4 days after removing systematic bias of the model outputs. Tian and Martinez (2012a,b) employed NCEP Global Forecast System (GFS) retrospective forecast (reforecast) data to generate 1-15 day probabilistic daily ET_0 forecast and then statistically downscaled the forecasts using different analog-based approaches in the southeastern United States. The results showed that most of the forecasts were skillful in the first 5 lead days. Tian and Martinez (2014) also generated the forecasts with the second-generation NCEP GFS reforecast dataset, which was operationally available from 2012 and included a complete set of meteorological data for the ET_0 estimation, with 11 ensemble members and a spatial resolution of 100 km. Compared with the previous studies, Tian and Martinez (2014) improved the skill of the probabilistic ET_0 forecasts as well as the performance of the soil water deficit estimation for irrigation scheduling in the first 5 lead days, due to the availability of a complete meteorological dataset produced by a more advanced NWP model at higher spatial resolution. Nonetheless, all the NWP-based ET_0 studies focused on either a single model, or a specific climate region, or a single aspect of forecast performance (Pelosi et al., 2016; Perera et al., 2014; Tian and Martinez, 2012a,b; Tian and Martinez, 2014; Perera et al., 2014). It is still lacking a comprehensive assessment of medium-range ET_0 forecasts based on multiple global NWP models over diverse climate regions.

Forecasts provided by the THORPEX Interactive Grand Global Ensemble (TIGGE) project provide an opportunity to produce next-generation medium-range ET_0 forecasting, given its multi-model ensemble feature, real-time accessibility, complete coverage in space and time, and fully archived near-surface variables (Swinbank et al., 2016). TIGGE is an unprecedented effort to

accelerate improvements in the accuracy of 1-day to 2-week high-impact weather forecasts. The TIGGE archive contains medium-range forecasts from nine operational, global ensembles, produced by the most important forecast systems, including the NCEP system. More importantly, TIGGE databases enable combining multiple model ensembles as an alternative for reducing ET_0 forecast errors, which it has not been explored in previous studies.

Hagedorn et al. (2012) found that a multi-model ensemble combining all nine models from the TIGGE archive did not outperform the best single model for temperature forecasts. However, a reduced multi-model system, consisting of only the best four model systems, the Canadian Meteorological Centre (CMC), the NCEP, the European Centre for Medium-Range Weather Forecasts (ECMWF), and the United Kingdom Meteorological office model (UKMO) showed improved performance, with the ECMWF model contributing the most to the added benefits of the multi-model and the CMC practically adding a negligible contribution. The ECMWF and UKMO models are accessible in real-time and have similar overall levels of skill (Buizza, 2014; Buizza et al., 2005; Johnson and Swinbank, 2009; Matsueda and Endo, 2011; Titley et al., 2008). Therefore, the multi-model ensembles through combining NCEP models with ECMWF and UKMO models would potentially improve ET_0 forecast performance.

This study aims to produce and evaluate deterministic and probabilistic ET_0 forecasts from both single and multi-model combinations of ECMWF, NCEP and UKMO forecasts from the TIGGE dataset. This study also evaluates the effects of different ensemble sizes on ET_0 forecast performance, as well as the impact of the individual weather forecast variable on ET_0 forecasts. Further, the study focused on the summer season and includes sites distributed over all climate regions in the contiguous United States (CONUS). This is the first study explicitly examining both probabilistic and deterministic forecasts from leading global NWP models from the TIGGE dataset and the first exploring the potential of using multi-model forecasts for improving medium-range ET_0 predictions.

2 DATASETS AND METHODS

2.1 Measurement dataset

In this study, daily measurements from 101 quality-controlled U.S. Climate Reference Network (USCRN) weather stations were used as the observational reference. As shown in Figure

II.1, these stations are distributed over nine climatologically consistent regions in CONUS divided by scientists of the National Centers for Environmental Information (NCEI) (Karl and Koss, 1984). We used USCRN stations, instead of the agricultural weather stations, for facilitating the comparisons among forecasting methods in different climate regions. The agricultural weather stations have a modified near-surface boundary layer compared to the surrounding landscape, providing a source of bias between observations and forecasts, whereas, the USCRN stations have been deployed in the locations that are representative of the climate of the region, and not heavily influenced by unique local factors (NOAA/NESDIS, 2003).

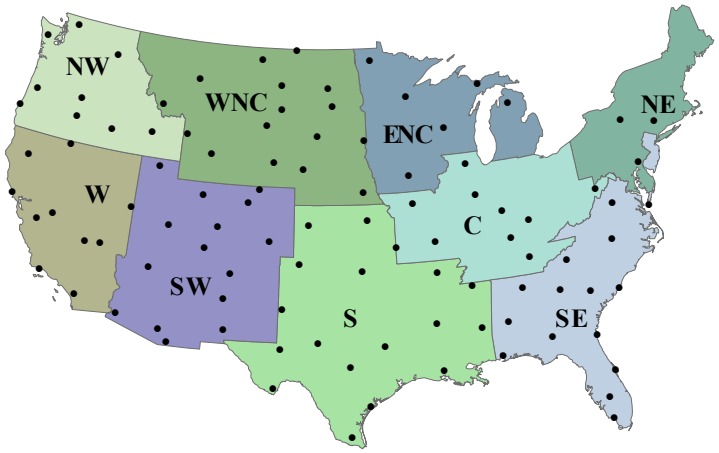

Figure II.1. U.S. climate regions: NW (North West), WNC (West North Central), ENC (East North Central), NE (North East), C (Central), SE (South East), S (South), SW (South West), W (West). The circles represent the sampled USCRN stations in the experiment.

High-quality observational climate datasets with thorough quality control are important for accurate ET_0 estimations. The USCRN meets the highest standards for instrument exposure (Menne et al., 2010) and is consistent in time and by sensor type (Leeper et al., 2015). The quality controls of the USCRN observations are periodically performed by NOAA's Atmospheric Turbulence and Diffusion Division to calibrate the sensors against the National Institute of Standards and Technology traceable standards that are re-certified annually. Automated USCRN field observations are monitored by the National Centers for Environmental Information, where indications of missing messages or questionable observation data are identified and addressed. Previous studies have shown that the quality-controlled USCRN observations such as solar radiation performed very well compared to the other observational references at daily timescale (e.g. Wang et al. 2012). The quality controlled USCRN observations would, therefore, be sufficient for providing high-quality ET_0 estimations. The reader can refer to the NCEI portal (https://www.ncdc.noaa.gov/crn/) for detailed information related to the USCRN data and data collection process.

The dataset used in this study comprised the period between May 1 and September 9 from 2014 to 2016. The retrieved data included daily observations of minimum and maximum temperature (T_{min} and T_{max} [ºC], respectively), minimum and maximum relative humidity (RH_{min} and RH_{max} [%]), and time-integrated surface incoming solar radiation (R_s [MJ m^{-2}]). Five minute-average wind speed values (u [m s^{-1}]) at a height of 1.5 m were also retrieved and aggregated to daily values. The latitude [º], longitude [º], and altitude [m] of the involved stations were also included in the original dataset.

Data of stations with a considerable number of missing records in a year were excluded from the dataset: NV_Baker_5_W data of the year 2014; TX_Austin_33_NW data of the year 2015; FL_Sebring_23_SSE and CO_Montrose_11_EN data of the year 2016. Days accounting for missing measurements in the remaining stations were removed from the analyses; these days represented a negligible portion of the total period.

2.2 Forecast dataset

The basic forecast datasets used in this study included perturbed ensemble forecasts issued by the ECMWF, NCEP, and UKMO models. These forecasts are freely available in the TIGGE data portal (see http://apps.ecmwf.int/datasets/data/tigge), over the period from October 2006 to the present. Table II.1 provides some important features of these three forecast systems.

Table II.1. Features of the involved forecast systems.

Centre	Resolution	Size of the perturbed ensemble	Forecast Length (h)
ECMWF	0.5° × 0.5°	50	360
NCEP	1.0° × 1.0°	20	384
UKMO*	0.4° × 0.3°	11	174

* Notice that the configuration of the UKMO forecast system was changed to the current setting from July, 2014.

The retrieved variables included: 2-m air maximum and minimum temperatures (T_{max} and T_{min} [ºK]), 2-m dew point temperature (T_d [ºK]), time-integrated surface net solar radiation (R_{ns} [J m^{-2}]), and the 10-m u and v components of the wind vector U [m s^{-1}]. The retrieved datasets comprised the 00 UTC (Coordinated Universal Time) perturbed forecasts issued between May 1 and August 31 from 2014 to 2016. ECMWF and NCEP datasets had a maximum lead time of 10 days, while UKMO consisted of a maximum lead time of 7 days. As the verification data USCRN

is recorded every day at local time, the measurements for a given day *d* were matched with the forecast values comprised between the +6 UTC hours of the day *d* and the +6 UTC hours of the day *d* + 1. All the retrieved forecasts were interpolated to the same 0.5° × 0.5° grid via the TIGGE data portal.

There were missing values in the retrieved NCEP and UKMO forecasts. The NCEP forecasts had 23 days missing in 2016 and four days missing in 2015. The UKMO forecasts of 2014 were available only up to July 15. The days with missing values were all discarded. The retrieved datasets included nine large grib2-format data files, one per system per year. Each grib2 file included perturbed forecast members for the grid cells between 26° and 49° North and 74° and 124° West. Forecasts from grib2 data files were interpolated to the USCRN stations using the nearest neighbor approach and were converted into manageable csv-format data files, one per station, per system and year. Wgrib2, a NCEP utility specifically designed to manipulate grib2 files, was used for processing the grib2 data. All the codes for data manipulation, analysis, and representation were scripted in R (R Core Team, 2014).

2.3 ET_0 estimation

ET_0 estimates based on USCRN measurements and NWP forecasts were computed using the FAO-56 PM (Allen et al., 1998, Allen et al., 2005) equation, which was available in the Evapotranspiration R package (Guo et al., 2016). The FAO-56 PM equation is recommended as the standard method for estimating ET_0. It applies energy balance and mass transfer principles to estimate the evapotranspiration from a uniform grass reference surface. Specific parameters are employed to model the surface and aerodynamic resistance from vegetation (Allen et al., 1998). The PM equation is expressed as follows:

$$ET_0 = \frac{0.408\Delta(R_n - G) + \gamma \frac{900}{T+273} u_2 (e_s - e_a)}{\Delta + \gamma(1 + 0.34u_2)} \quad \text{(II.1)}$$

where ET_0 is the daily reference evapotranspiration [mmday^{-1}], R_n is the net surface radiation at the crop surface [MJm^{-2}day^{-1}], G is the soil heat flux density [MJm^{-2}day^{-1}], T is the daily mean air temperature at 2 m height [°C], u_2 is the wind speed at 2 m height [m s^{-1}], e_s is the saturation vapor pressure [kPa], e_a is the actual vapor pressure [kPa], Δ is the slope of the vapor pressure curve [kPa°C^{-1}] and γ is the psychometric constant [kPa°C^{-1}].

The input variables for the FAO-56 PM equation included: minimum and maximum temperatures [°C], minimum and maximum relative humidity [%], surface incoming solar radiation [MJ m^{-2}], and the wind speed [m s^{-1}]. Net solar radiation forecast should be transformed into incoming solar radiations using the expression $R_s = R_n/(1-\alpha)$, where $\alpha = 0.23$ is the albedo of the "reference crop" evaporative surface. Minimum and maximum relative humidity forecasts were estimated using the maximum and minimum temperature and the dew point temperature forecasts, following Lawrence (2005). Daily 10-m wind speed forecasts were adjusted to the same height of the USCRN measurements assuming a vertical log wind profile (Allen et al. 1998). Details about the implementation of the routine for computing ET_0 can be found in (Guo et al., 2016) or as part of the R package (Guo and Westra, 2017).

2.4 Bias correction of ET_0 forecasts

To investigate the effect of post-processing on ET_0 forecasts, a simple deterministic correction (a.k.a. calibration) was conducted by fitting a linear regression between observations and ensemble mean forecasts from a training dataset and applying the regression to the current mean forecast value. The correction factor, given by the difference between the regression adjusted forecast and the raw ensemble mean forecast, was added to all ensemble members of the raw forecast in a way that the ensemble distribution was shifted. The training data included the forecast-observation pairs corresponding to the 30 days before the forecast initial day. The calibration procedure was applied at all stations and lead times.

2.5 Forecast evaluation

Hereafter, we refer to the measurement-based ET_0 estimations as "observed ET_0" or "observations", to ET_0 forecasts before calibration as "raw" forecasts, and to the ET_0 forecasts after bias correction either as "bias-corrected" or "calibrated" forecasts. Both raw and calibrated ET_0 forecasts were evaluated against USCRN based observations for each lead time, year, station and system. The different statistics were averaged over each NCEI climate region (Figure II.1) to show the performance of each forecast in climatologically different regions. The forecasts were assessed using both deterministic and probabilistic metrics. Deterministic metrics were used for evaluating the ensemble mean forecasts; probabilistic metrics were used to evaluate probabilistic

forecasts, which were converted from the ensemble forecasts. Detailed information for each forecast metric is described below.

2.5.1 Deterministic forecast metrics

For the deterministic metrics, we used the mean error (ME) and the root mean square error (RMSE), which are among the most commonly reported measures of agreement between forecasts and observations. ME provides an estimate of the model bias while RMSE is an accuracy measuring criteria. For unbiased models, RMSE is an estimator of the square root of the ensemble variance.

Let $\bar{f}_{l_r,d_l,t}$ represent the mean of the forecast variable at location l_r $(l_r = 1{:}\,N_{l_r})$ and day d_l $(d_l = 1{:}\,N_{d_l})$, with a lead time *t*. N_{l_r} is the number of locations belonging to a region *r*, while N_{d_l} is the total number of available forecast days for a specific location. Let o_{l_r,d_l} denote the observed variable at the corresponding location and day. The ME and RMSE for the specific region *r* and lead time *t*, $\mathrm{ME}_{r,t}$ and $\mathrm{RMSE}_{r,t}$, respectively, are then computed as:

$$\mathrm{ME}_{r,t} = \frac{1}{N}\sum_{l_r=1}^{N_{l_r}}\sum_{d_l=1}^{N_{d_l}}\left(\bar{f}_{l_r,d_l,t} - o_{l_r,d_l}\right) \quad \text{(II.2)}$$

$$\mathrm{RMSE}_{r,t} = \sqrt{\left(\frac{1}{N}\sum_{l_r=1}^{N_{l_r}}\sum_{d_l=1}^{N_{d_l}}\left(\bar{f}_{l_r,d_l,t} - o_{l_r,d_l}\right)^2\right)} \quad \text{(II.3)}$$

where *N* is the total number of pairs of forecasts and observations in a specific region.

2.5.2 Probabilistic forecast metrics

The skill of the probabilistic forecast was evaluated using the Brier Skill Score (BSS) associated with the tercile events of the ensemble forecasts (upper or 1st, middle or 2nd, and lower or 3rd terciles). Let $p_{l_r,d_l,t}$ represent the forecast probability of the considered event occurring at location l_r and day d_l with lead time *t*. Let o_{l_r,d_l} be equal to 1 if the event occurs at the specific location and day and 0 otherwise. Similarly to RMSE in the deterministic case, the Brier Score (BS) measures the mean squared probability error (Murphy, 1973) of the forecast associated with a given threshold value (or event). The BS for the specific region *r* with lead time *t* ($\mathrm{BS}_{r,t}$) is then calculated as follows:

$$\mathrm{BS}_{r,t} = \frac{1}{N}\sum_{l_r=1}^{N_{l_r}}\sum_{d_l=1}^{N_{d_l}}\left(p_{l_r,d_l,t} - o_{l_r,d_l}\right)^2 \quad \text{(II.4)}$$

The corresponding Brier Skill Score ($\mathrm{BSS}_{r,t}$) then measures the improvement of the probabilistic forecast relative to a reference forecast, usually called the sample climatological distribution, or the sample climatology (Wilks, 2011):

$$\mathrm{BSS}_{r,t} = 1 - \frac{\mathrm{BS}_{r,t}}{\mathrm{BS}_{\mathrm{clim}_r}} \tag{II.5}$$

where $\mathrm{BS}_{\mathrm{clim}_r}$ refers to the Brier Scores of the sample climatology, which is defined in this study as a function of the relative frequencies of the *N* observations o_{l_r,d_l} in the verification data set (Wilks, 2010):

$$\mathrm{BS}_{\mathrm{clim}_r} = \bar{o}_{l_r,d_l}\left(1 - \bar{o}_{l_r,d_l}\right) \tag{II.6}$$

where $\bar{o}_{l_r,d_l}$ is the event relative frequency within the *N*-member sample of observations. The BSS ranges from -∞ to 1 and values of BSS equal to 1 indicate perfect skill.

Since, in this study, the BSS was used to evaluate the skill associated with the tercile events of the ensemble forecasts, the event relative frequency $\bar{o}_{l_r,d_l}$ is, in all cases, constant and equal to $0.3\overline{3}$ and $\mathrm{BS}_{\mathrm{clim}_r} = 0.2\overline{2}$, and, consequently, the BSS values for the different regions and lead times are inversely proportional to the corresponding BS values.

Binary events highlight only one aspect of the forecast. The Continuous Rank Probability Score (CRPS), which is recommended to obtain a broader overall view of performance (Hersbach, 2000), was also used to evaluate the probabilistic forecast performance. The CRPS is precisely the integral of the Brier scores at all possible threshold values for the continuous predictand (Gneiting et al., 2005; Hersbach, 2000). It measures the integrated square, by all possible threshold values *h*, of difference between the cumulative distribution function (cdf) of the forecast variable, F^f, and the corresponding cdf of the observed variable, F^o. Following Gneiting et al. (2005), the CRPS of an ensemble forecast for lead time *t*, location l_r and day d_l ($\mathrm{crps}_{l_r,d_l,t}$) is computed as

$$\mathrm{crps}_{l_r,d_l,t} = \int_{-\infty}^{\infty}\left(F^f_{l_r,d_l,t}(h) - F^o_{l_r,d_l}(h)\right)^2 dh \tag{II.7}$$

where $F^o_{l_r,d_l}(h) = H\left(h - o_{l_r,d_l}\right)$, *H* being Heaviside function, which takes the value 0 when $h < o_{l_r,d_l}$ and value 1 otherwise. The aggregated continuous rank probability score for a region *r* and lead time *t*, $\mathrm{CRPS}_{r,t}$ is then computed as:

$$\mathrm{CRPS}_{r,t} = \frac{1}{N}\sum_{l_r=1}^{N_{l_r}}\sum_{d_l=1}^{N_{d_l}}\mathrm{crps}_{l_r,d_l,t} \tag{II.8}$$

The CRPS variates between 0 and +∞; smaller values indicate better performances.

Reliability diagrams, as a measure of systematic and conditional bias, were also computed to investigate the reliability of the forecasts. The reliability diagram plots the observed frequency of an event (defined by the threshold h) against its forecasted probability. The range of forecast probabilities is divided into k bins, then, on the x-axis, we plot the average probability of the forecasts that fall in the k-th bin, while on the y-axis, the fraction of the corresponding observations that are below the threshold is plotted. Perfect reliability is achieved along the 45° diagonal line on the reliability diagram when the observed frequency of the given event within each bin equals the average of the corresponding forecast probabilities. The deviation from the diagonal gives the conditional bias. On the reliability diagram, it is also possible to show the sharpness of the forecast, which is a measure of the forecast confidence, using a histogram representing the frequency of forecasts in each probability bin. Sharper forecasts mean more concentrated frequency distributions of the ensemble forecasts. Sharper forecasts usually indicate better forecasts if they have good reliability or calibration (Gneiting et al., 2007).

A bootstrapping analysis was also included in the reliability diagrams to assess the uncertainty of the sampled pairs of points. Random sampling was performed 1000 times with replacement in a standard way. Error bars accounting for the 5 and 95 percent of the distribution of the values were indicated in the reliability diagrams.

2.6 Inter-comparison of forecast schemes

The study compared the performance of single ET_0 raw forecasts of the ECMWF, NCEP, and UKMO, as well as the multi-model ET_0 forecasts arising from the simple combination of the three ensemble forecasts (hereinafter referred to as ECMWF-NCEP-UKMO) and the combination of the two ensemble forecasts (the ECMWF and UKMO ensembles, hereinafter referred to as ECMWF-UKMO). The differences between the performance of calibrated forecasts and raw forecasts for each system and the multi-model ensemble system were also investigated. All the comparisons were conducted over the same period between May 31 and August 31 from 2014 to 2016, with a bootstrapping analysis (as specified below) applied to assess if the differences were statistically significant.

When assessing the difference, the weather forecast data should be treated as paired since it accounts for the error correlation between two samples (Hamill, 1999). When performing the bootstrapping analysis, each of the system forecasts and observations was simultaneously sampled

from the same randomly chosen dates. The bootstrapping analysis relied on 1000 random sampling with replacement. The differences were considered significant if at least 95% of the bootstrapped differences had the same sign (positive or negative).

We also examined the effects of different number of ECMWF perturbed members on the performance of ET_0 forecasts. Considering the large ensemble size of this NWP system, this is useful from both a theoretical and practical point of view. For example, it is important to know how a NWP system behaves when it has the same number of ensemble members as the other systems. Also, a reduced number of ensemble members is more efficient in operational schemes where computation is time-consuming. This study compared the performance of a NWP forecast with a different number of randomly sampled ensemble members, applied to both raw and calibrated forecasts.

2.7 Impact of the individual weather forecast variable on ET_0 forecasts

We also evaluated the forecast performance for individual weather variables and investigated how the forecast uncertainty of the individual weather variable affected the ET_0 forecasts. The analysis was performed by comparing the error distributions of the raw forecasts for four individual weather variables T, RH, Rs and u, with the corresponding error distributions of the raw ET_0 forecasts. The ET_0 forecasts were calculated using one forecast variable and three observed variables, i.e. by replacing one observation at a time with one forecast, so that the impacts of a single forecast variable on ET_0 forecasts can be separated. When perturbing T and RH, we simultaneously replaced the maximum and minimum observed values by the corresponding forecasts.

While we assessed the impact of individual weather variables on ET_0 forecast skill, we did not evaluate the skill of individual weather variables. The reasons were twofold. First, the scope of this paper is to produce and assess multi-model ET_0 forecasts, which includes multiple experiments and procedures that provide ample information for understanding and improving ET_0 forecasts. Second, while the bias of the ET_0 forecast is caused by individual weather variables, bias correcting calculated ET_0 is computationally more efficient and addresses the biases from both, individual weather variables as well as the correlations among variables. Previous studies

have found that bias corrected ET_0 forecasts performed better than any individual input variables (e.g. Lewis et al., 2014).

3 RESULTS AND DISCUSSION

3.1 Variability of ET_0 observations

The ET_0 observations based on the USCRN measurements show diverse distributions over the NCEI climate regions (Figure II.2). The mean values approximately vary between a minimum of 3.2 and a maximum of 6.3 mm day^{-1}, while the standard deviation ranged from 1 to 1.8 mm day^{-1}. Both mean and variance of ET_0 forecasts are consistently higher in southern regions, which comprise coastal and continental stations and a complex orography, like the W and SW regions. June and July ET_0 values tend to be higher than those in May and August, while May and June ET_0 values show higher variability.

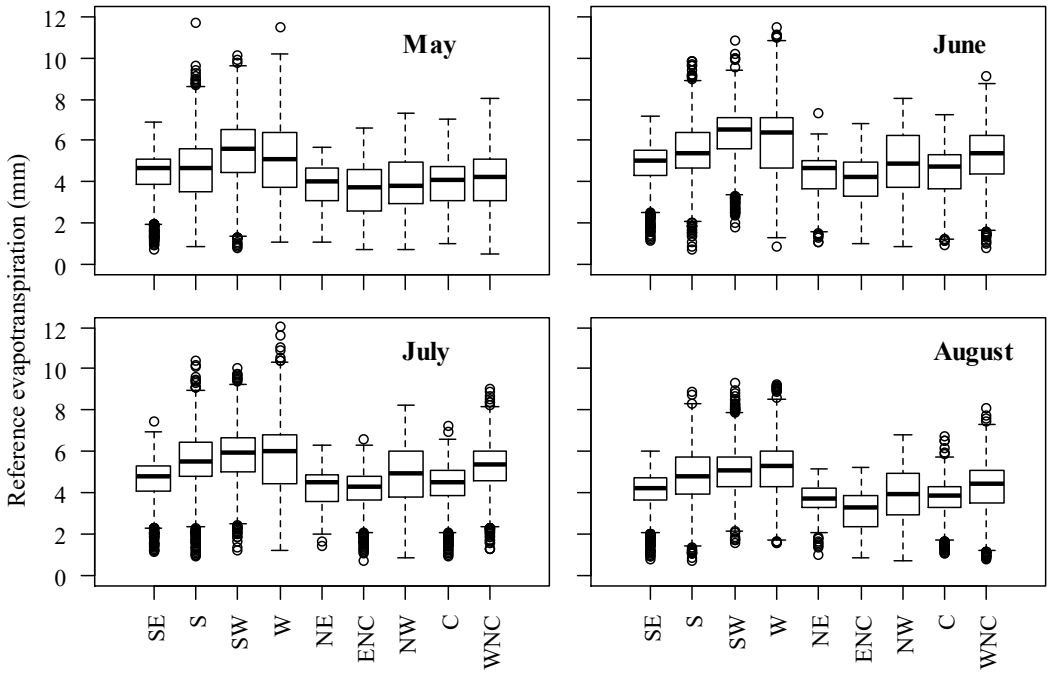

Figure II.2. Box-whisker plots of the measurement-based daily ET_0 estimates for each climate region in May, June, July and August, from 2014 to 2016.

3.2 Performance of raw ET_0 forecasts

Here we provide a comparative assessment of the performances of the single- and multi-model raw ET_0 forecasts across regions and for different lead times. Figure II.3 shows the ME and RMSE for the different system forecasts.

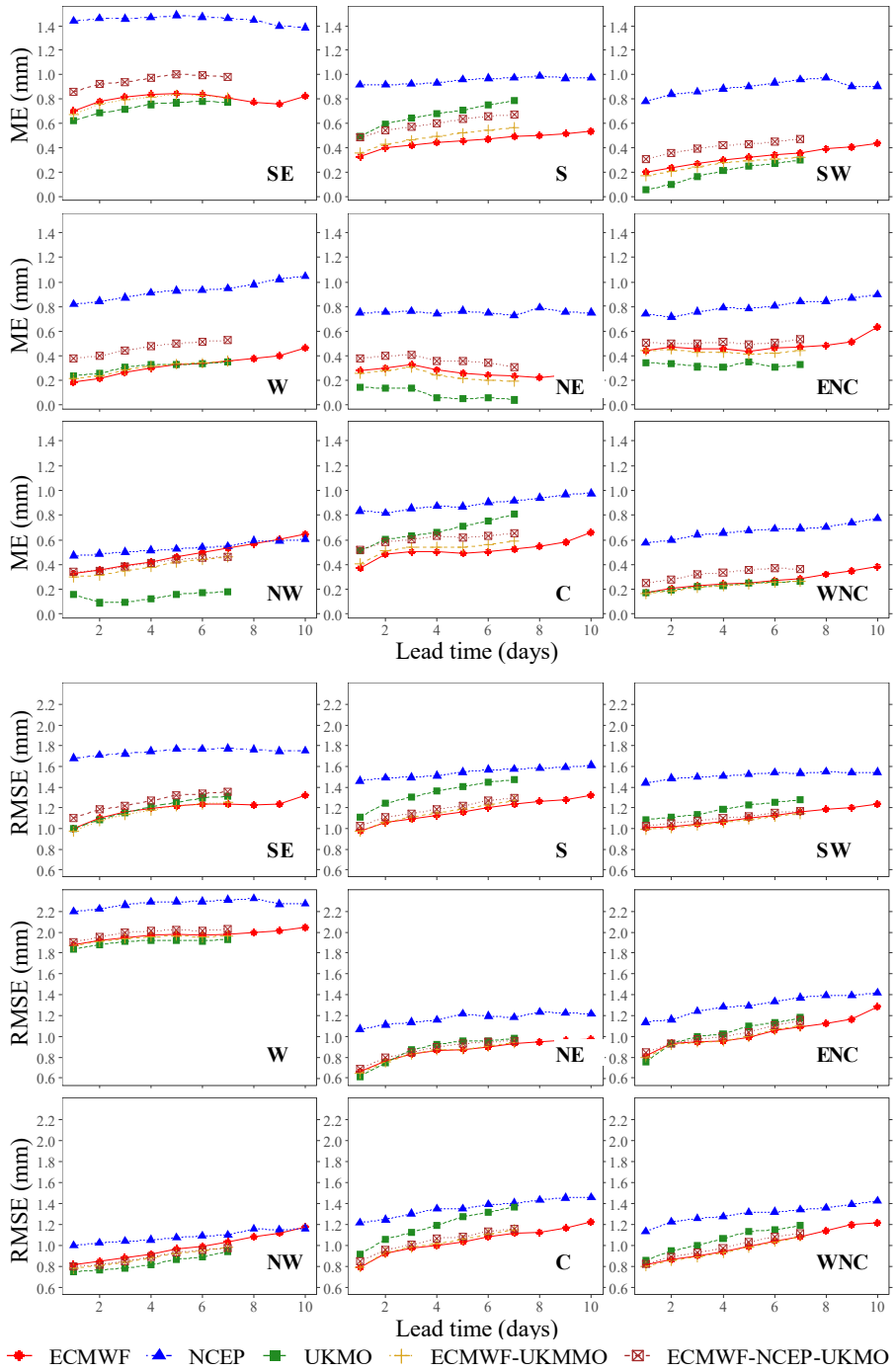

Figure II.3. ME and RMSE of the raw ECMWF, NCEP, UKMO, ECMWF-UKMO, and ECMWF-NCEP-UKMO ET_0 forecasts for each climate region as a function of lead time.

All systems tend to over-predict ET_0 observations for different lead times and regions. The raw NCEP forecasts are more biased and less accurate than the raw ECMWF and UKMO forecasts. The raw ECMWF forecasts show slightly higher ME but much lower RMSE than the UKMO forecasts, for longer lead times. The performances of the raw ET_0 forecasts also vary among the different regions. All systems perform sensibly better in the northern regions than in southern regions and the west region. The performance is poor in the W region, which is characterized by a very complex landscape. The simplified terrain heights of the forecast model may probably affect the quality of the forecasts in regions like this (Hagedorn et al., 2008).

Figure II.4 shows the BSS of the raw ECMWF forecasts for the three terciles of the ensemble forecasts, as well as the differences between the BSS of the NCEP, UKMO, ECMWF-UKMO and ECMWF-NCEP-UKMO systems and the BSS of the ECMWF system. The results indicate that, for probabilistic ET_0 forecasts, the BSS is generally better for the lower tercile event, compared with the upper and middle tercile events. Most of the middle tercile forecasts show low BSS and reliability since the changes or shifts in the middle tercile are not as sizeable as in the lower or upper terciles, as found and discussed in previous studies (e.g Barnston et al., 2003; Van Den Dool and Toth, 1991). In Figure II.5 and II.6, we compare the reliability diagrams related to the different ET_0 systems forecasts for the upper and lower tercile events, respectively, considering for the comparison, in the former case, different regions at the same lead time (i.e. three days) and, in the latter case, different lead times at the same region (i.e. the S region).

The ensemble forecasts show an over-forecasting bias when predicting ET_0 values in the upper tercile (Figure II.5) and an under-forecasting bias for the lower tercile event (Figure II.6). ET_0 forecasts for the first tercile events are generally more reliable and sharper than for the third tercile events. The reliability for short and middle lead times is similar in most experiments (see Figure II.6). In general, all the systems' forecasts, particularly the NCEP forecasts, seem fairly sharp.

The results, reported in Figures II.4 to II.6, indicate that NCEP ensembles are considerably less skillful and reliable than any other ensemble. The ECMWF and UKMO performances are comparable: the ECMWF forecasts are on average more skillful and reliable, although UKMO performs similarly or slightly better in experiments for the northern regions and the west region.

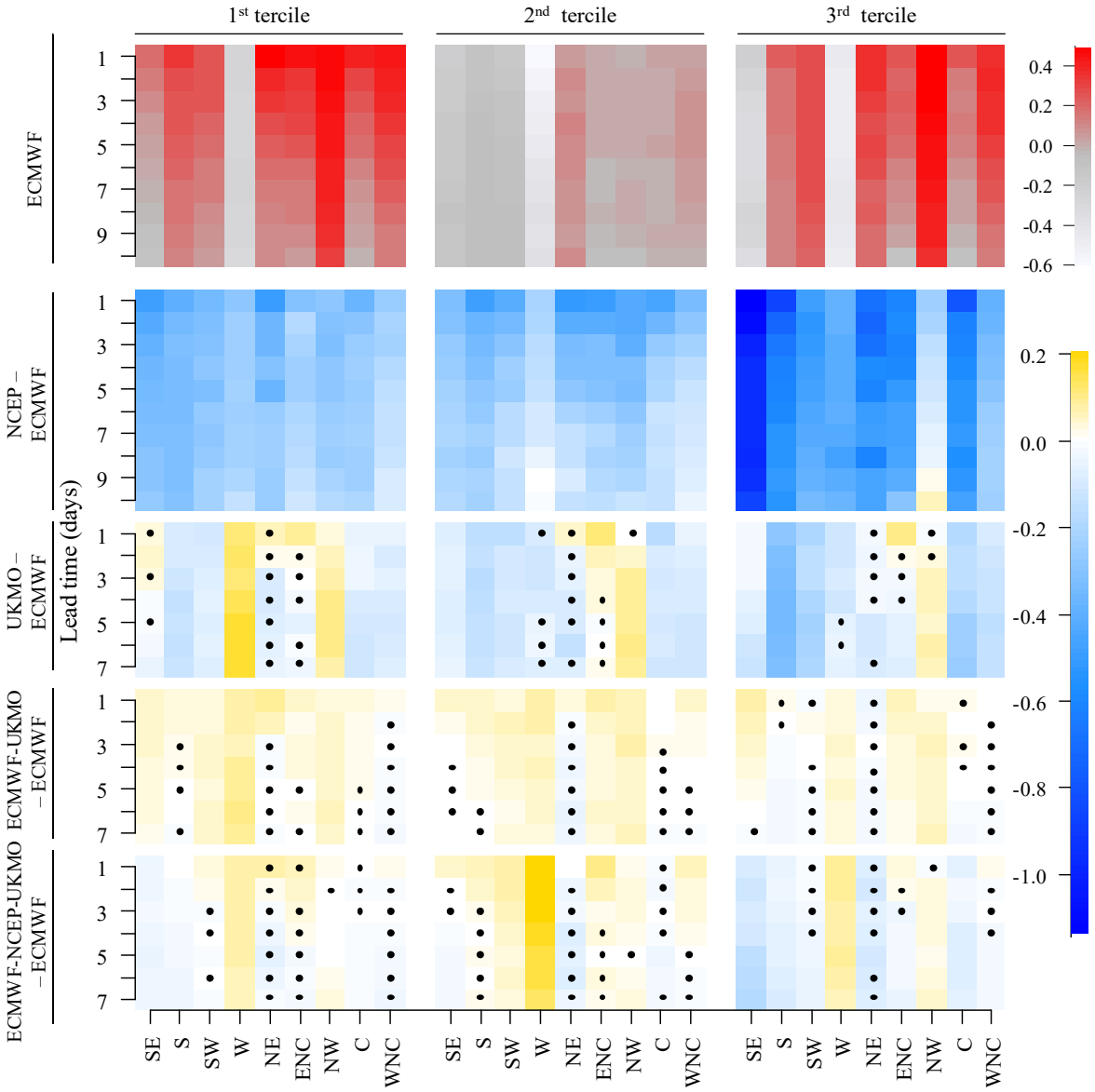

Figure. II.4. BSS or BSS difference of the ET_0 forecasts. The top row is the BSS of the ECMWF forecasts for each tercile. The second to fifth rows represent the differences between BSSs of the NCEP, UKMO, ECMWF-UKMO, and ECMWF-NCEP-UKMO forecasts and the ECMWF forecasts. The black dots indicate the instances where the differences are not significant.

As a result, the ECMWF-UKMO forecasts guarantee equal or higher BSS compared with ECMWF and ECMWF-NCEP-UKMO in most cases. The overall statistics for the multi-model ECMWF-NCEP-UKMO are presumably affected by the poor performance of NCEP. Hagedorn et al. (2008) and Mathiensen and Kleissl (2011) showed that original ECMWF forecasts of temperature and radiation, respectively, outperformed the corresponding NCEP forecasts in the continental U.S.

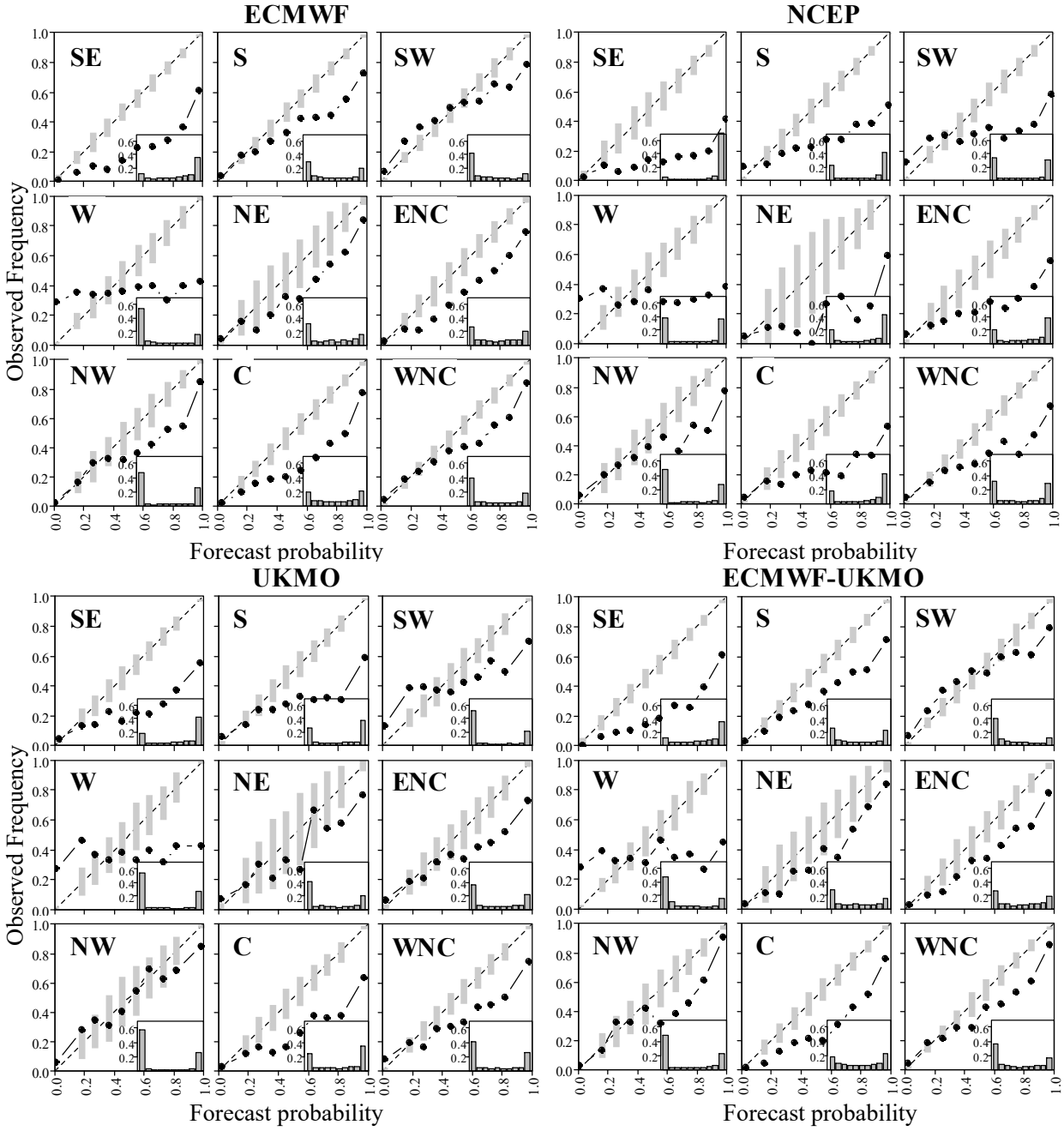

Figure II.5. Reliability diagrams of the raw ECMWF, NCEP, UKMO, and ECMWF-UKMO ET_0 forecast for the third tercile event at 3-day lead. Inner histograms show the relative frequency with which the event has been predicted for the different levels of probability.

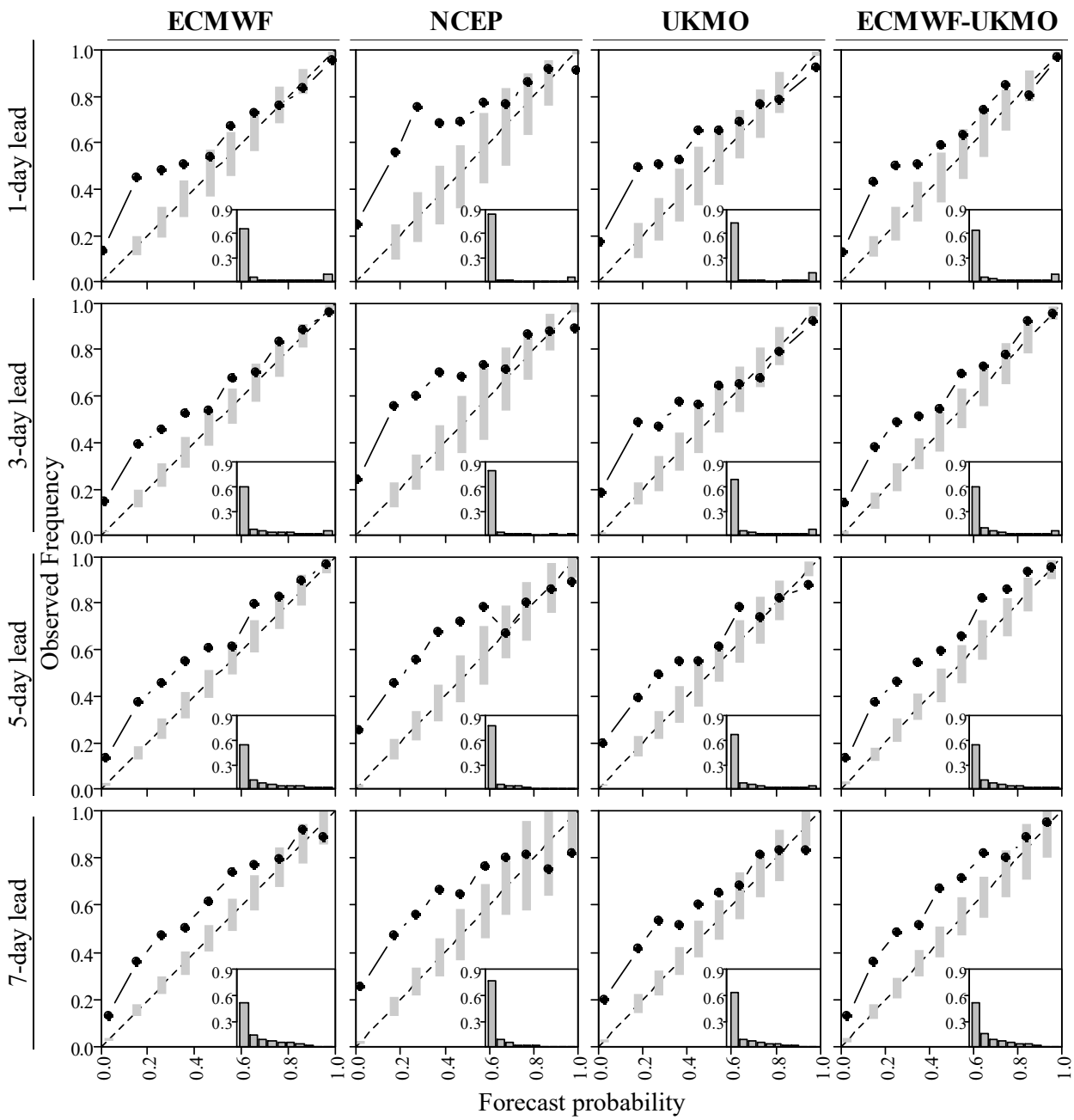

Figure II.6. Reliability diagrams of ECMWF, NCEP, UKMO, and ECMWF-UKMO based ET_0 forecasts for the first tercile event in the S (South) region at 1-, 3-, 5-, and 7-day lead times.

Table II.2 summarizes the median values of different metrics over all regions for the forecasts at 1- and 7-day leads. The median values using ECMWF-UKMO forecasts are consistently better than for ECMWF forecasts at 1-day lead, while they behave similarly at 7-day lead. The reason is that UKMO forecasts tend to degrade more than ECMWF forecasts over lead

time. The ECMWF-NCEP-UKMO multi-model practically provides no improvement compared to ECMWF-UKMO and ECMWF given the unfavorable performance of the NCEP forecasts.

Table II.2. Median values of different metrics for the raw forecasts at 1- and 7-day leads over all climate regions.

	ECMWF		NCEP		UKMO		ECMWF-UKMO		ECMWF-NCEP-UKMO	
	+1-day	+7-day	+1-day	+7-day	+1-day	+7-day	+1-day	+7-day	+1-day	+7-day
ME	0.327	0.471	0.779	0.916	0.238	0.330	0.297	0.439	0.378	0.524
RMSE	0.819	1.116	1.213	1.402	0.915	1.280	0.802	1.144	0.849	1.159
CRPS	0.534	0.631	0.876	0.882	0.586	0.780	0.499	0.651	0.501	0.664
BSS1	0.388	0.131	-0.013	-0.152	0.331	0.053	0.412	0.140	0.392	0.115
BSS2	-0.029	-0.038	-0.495	-0.247	-0.152	-0.138	0.022	-0.035	-0.011	-0.045
BSS3	0.245	0.105	-0.359	-0.380	0.064	0.012	0.257	0.099	0.249	0.037

3.3 Performance of calibrated forecasts

The regression calibrations substantially improve the probabilistic performance of the ensemble forecasts, especially regions and forecast systems with less skillful raw forecasts. As it is shown in Figure II.7, the differences in terms of BSS between calibrated forecasts and raw forecasts are mostly positive and significant for each model and all three terciles, with some exceptions for the middle tercile. The skill of the NCEP forecasts considerably improves after bias corrections, more than other forecast systems. The southern and western regions also gain more from calibrations than the northern regions, which were characterized by better performances for the raw forecasts. In particular, the W region goes from being worse to being the best through the simple bias correction. Meanwhile, the increments in terms of BSS for the upper tercile are at least twice as much as those for the lower event.

We further compare the BSS for the calibrated single- and multi-model forecast systems in Figure II.8. The NCEP forecasts are still less skillful than the ECMWF forecasts or any other systems' forecasts; the calibrated ECMWF-UKMO and ECMWF-NCEP-UKMO forecasts are significantly more skillful than the calibrated ECMWF forecasts in the experiments where ECMWF and UKMO showed similar performances, and mostly for short lead times and/or in the case of the W and NW regions. On the other hand, the multi-models are as skillful as ECMWF in the experiments where ECMWF outperformed UKMO.

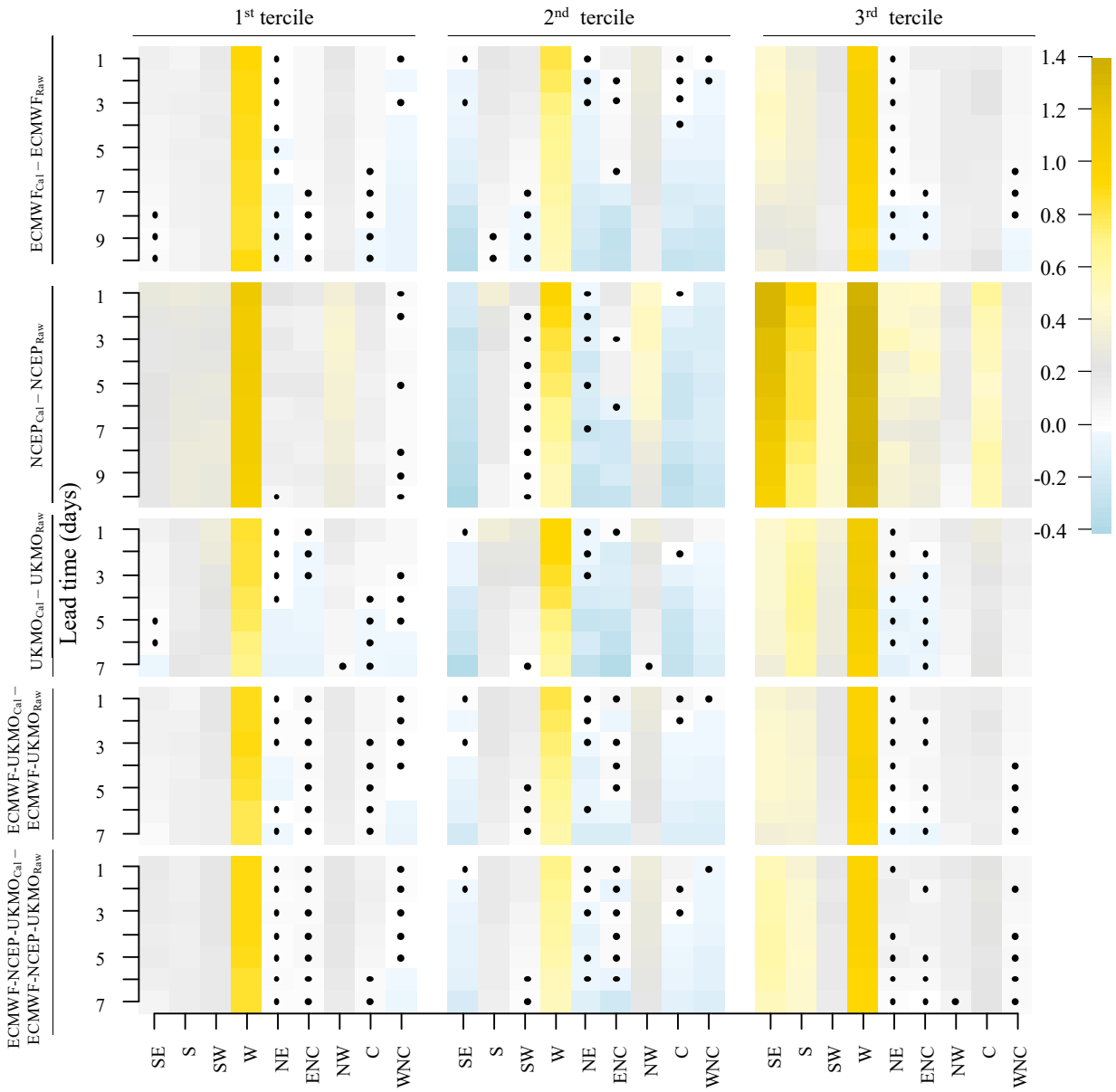

Figure II.7. Changes in the BSS for each calibrated ET_0 forecast scheme relative to the raw forecast. The solid dots indicate the differences are not significant. Raw and calibrated forecasts covered the common period between May 1 and August 31, 2014 to 2016.

The maximum improvements in terms of BSS from the multi-model systems represent a gain in forecast skill of less than one day compared with the single model systems. In addition to the BSS, the reliability of all the forecasts is also improved through the regression calibration process, with the third tercile events experiencing the highest improvements, as shown in Figure II.9. The calibrated ECMWF forecasts are more reliable but slightly less sharp than the NCEP and UKMO forecasts and are equally reliable and sharp as ECMWF-UKMO and ECMWF-NCEP-UKMO.

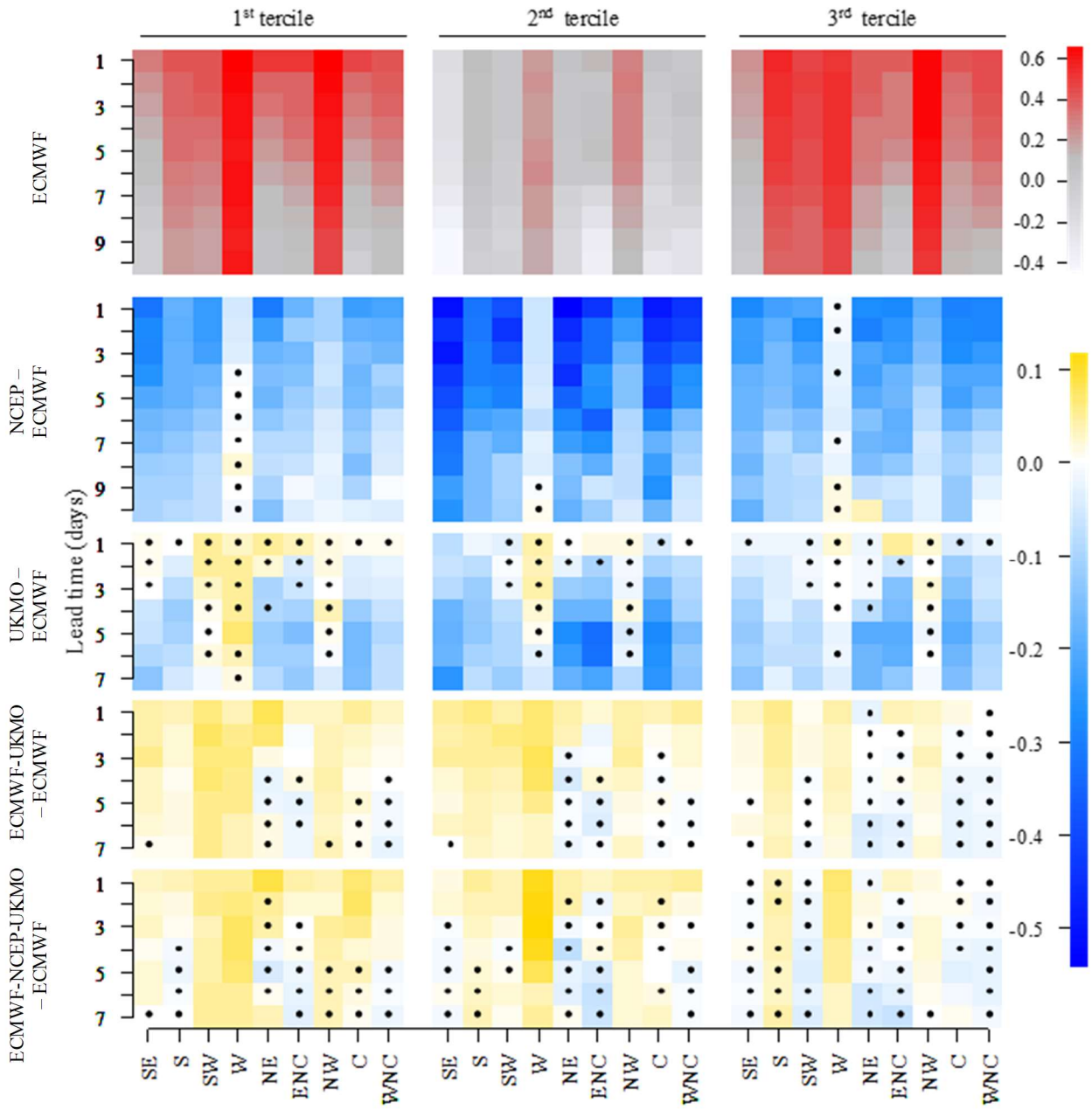

Figure II.8. As in Figure II.4, but for calibrated forecasts

It is worth noting that the reliability assessments might be affected by the sample size in regions such as W, NE, and NW, because the diagrams with relatively small sample size are prone to be noisy (e.g. Hamill, 1997). In Table II.3, we summarize the median values of the evaluated metrics for the calibrated forecasts at 1- and 7-day lead times. By comparing Tables 2 and 3, we find that the BSS in terciles, the ME, RMSE, and CRPS values are considerably improved after calibration.

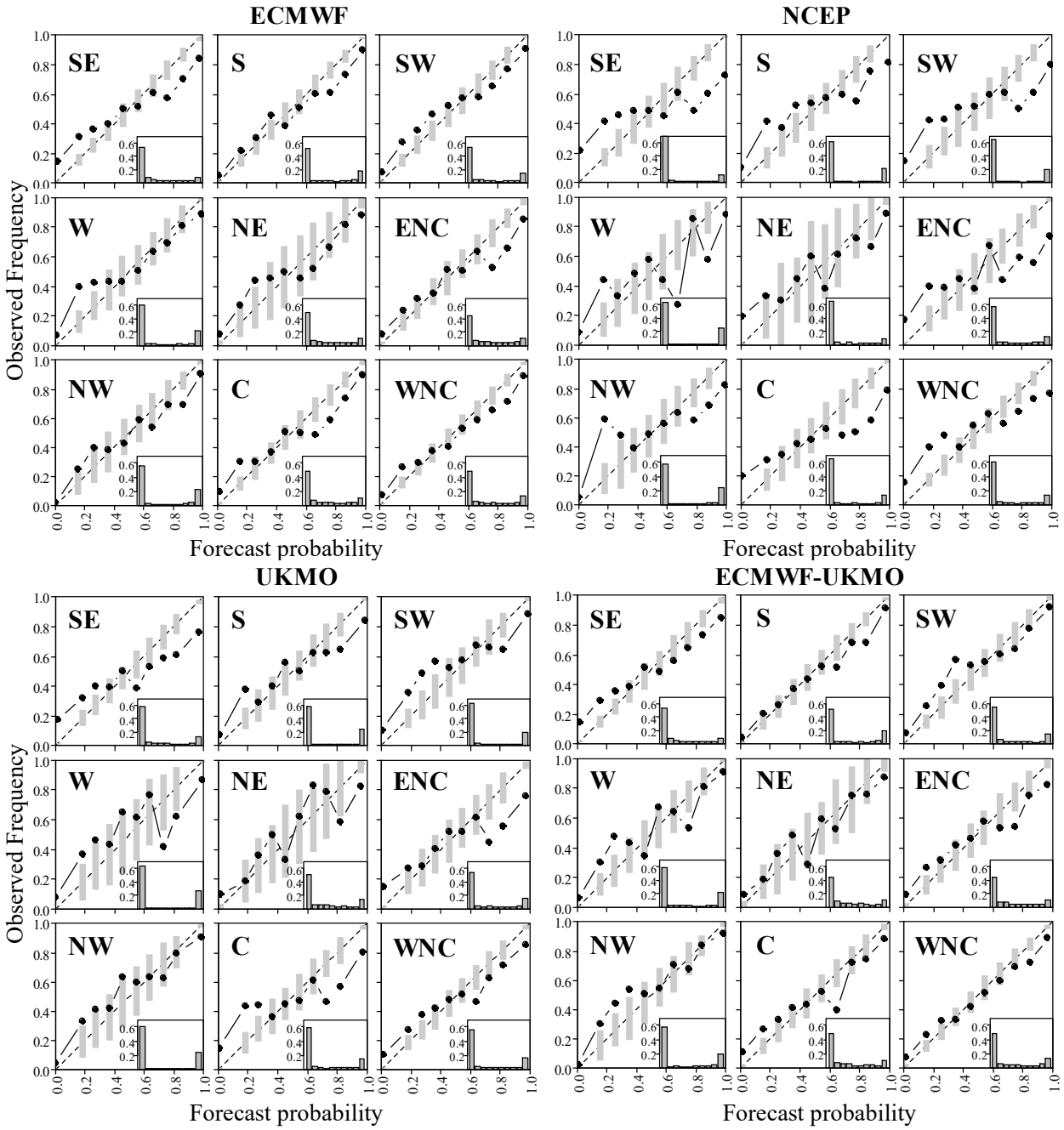

Figure II.9. Reliability diagrams of the calibrated ECMWF, NCEP, UKMO, and ECMWF-UKMO ET_0 forecasts for the third tercile event at 3-day lead.

The ECMWF-UKMO and ECMWF-NCEP-UKMO forecasts seem to be consistently better than the ECMWF forecasts at 1-day lead, but not at 7-day lead, since the forecast performance from the three systems is consistently low at 7-day lead. ECMWF-NCEP-UKMO is not recommended for implementation system as it does not bring additional improvements concerning either ECMWF or ECMWF-UKMO. It is worth noting that the RMSE values from

calibrated forecasts are mostly lower than the other NWP-based ET_0 forecasts reported in Australia using the ACCESS-G model (Perera et al., 2014) and Chile using MM5 (Silva et al., 2010). On the other hand, the BSS assessments for the SE region from the ECMWF and the multi-model forecasts are similar to those found by Tian and Martinez (2014) using NCEP based analog forecasts (Hamill and Whitaker, 2006; Hamill et al., 2006).

Table II.3. As in Table II.2, but for calibrated forecasts.

Stat.	ECMWF		NCEP		UKMO		ECMWF-UKMO		ECMWF-NCEP-UKMO	
	+1-day	+7-day	+1-day	+7-day	+1-day	+7-day	+1-day	+7-day	+1-day	+7-day
ME	0.038	0.061	0.026	0.045	0.014	0.066	0.036	0.070	0.024	0.055
RMSE	0.679	0.945	0.785	0.993	0.681	0.958	0.668	0.919	0.674	0.918
CRPS	0.402	0.557	0.539	0.619	0.416	0.599	0.386	0.538	0.383	0.539
BSS_{1st}	0.434	0.168	0.199	0.112	0.445	0.092	0.478	0.140	0.500	0.145
BSS_{2nd}	0.050	-0.093	-0.396	-0.247	0.040	-0.232	0.081	-0.098	0.086	-0.099
BSS_{3rd}	0.430	0.244	0.143	0.159	0.420	0.154	0.429	0.222	0.424	0.225

The NWP based ET_0 forecasts have been implemented in operational advisory systems to assist farmer's decision making (Chirico, et al. 2018). Studies like this provide a basis for an operational system of agronomic decision making. Nevertheless, the bias-correction approach used in this study is applied to the forecast at the USCRN stations, which are representative of the climate of the region but are not in agricultural settings. The bias corrected forecasts based on the USCRN settings will likely be biased warm and dry compared to agricultural settings. Therefore, a different bias-correction, i.e., based on agricultural stations, will be needed to take into account the microclimate in agricultural settings. This can be done by removing the air temperature and relative humidity errors as part of a preprocessing analysis (e.g. Lewis and Alan, 2017) or by directly bias correcting the calculated ET_0 forecasts (e.g. Tian et al., 2014) at agricultural weather stations.

3.4 The effect of ensemble size on ECMWF based ET_0 forecasts

Model systems with a greater number of ensemble members are expected to have better probabilistic forecasts (Ferro et al., 2008) as well as greater impacts on the multi-model predictions than model systems with fewer members (Hagedorn et al., 2012). This section analyzes in what extension the ECMWF based ET_0 forecasts are affected by the ensemble size.

Figure II.10 shows the BSS for the first and third tercile events using 10, 20, 30, 40 and 50 ensemble members of the raw and calibrated ECMWF forecasts.

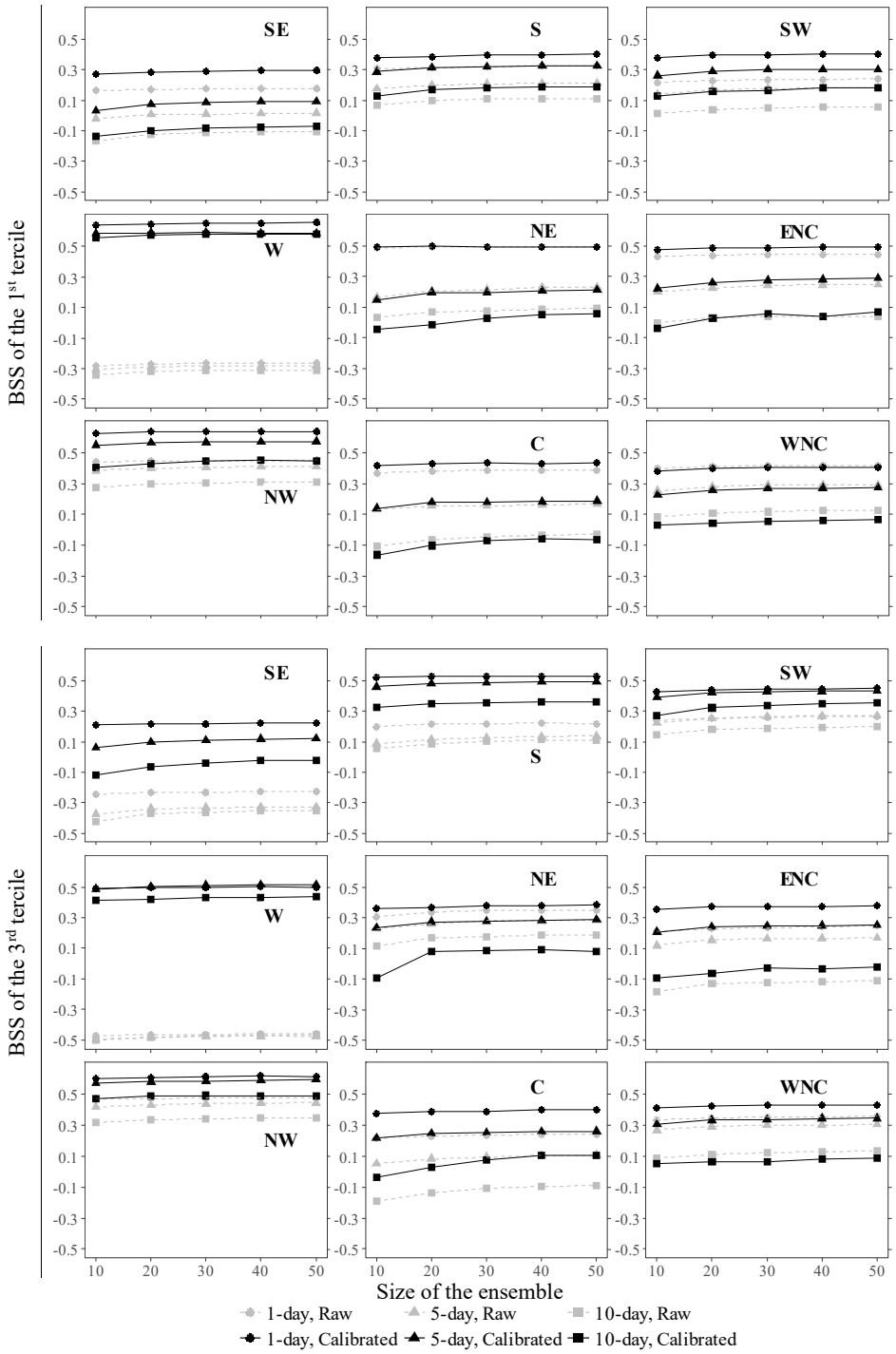

Figure II.10. BSS of the calibrated and raw ECMWF forecasts for ET_0 with 10, 20, 30, 40 and 50 ensemble members for the first and third tercile events over each climate region.

In agreement with Ferro et al. (2008), the larger the size of the ensemble, the higher the skill of the probabilistic forecasts. However, the changing rate of BSS is almost negligible for 1-day lead and increases just slightly with the increased lead time. For 5 and 10 day leads, an ensemble size of 20 and 30 members, respectively, seems to reach the same performance as with 50 members. The forecasting scheme (raw or bias corrected) scarcely affects the functional dependence between BSS or CRPS and ensemble size. The improvements for the third tercile events through calibration suggested that it was more efficient using a calibrated forecast with 10 ensemble members, than using a raw forecast model with full ensemble members, regardless of the lead times.

We further examine the median CRPS and BSS values of the 10-member ECMWF forecasts and the full-member UKMO forecasts for the first and third tercile events as a function of lead time (Figure II.11). The ECMWF forecasts perform better than the UKMO forecasts even with much fewer forecast members, except for the BSS of the raw forecasts for the first tercile. Notice that the ECMWF calibrated forecasts still improves the skill by approximately one day relative to the calibrated UKMO forecasts for 3-day lead or more.

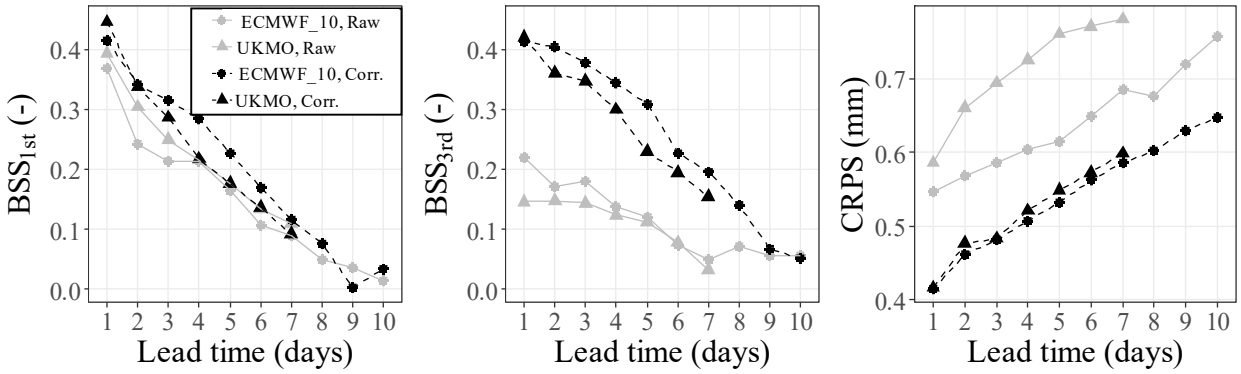

Figure II.11. Median values of the lower and upper tercile BSS (BSS_{1st} and BSS_{3rd}) and CRPS of the 10-member ECMWF forecasts for ET_0 and full-member UKMO forecasts for ET_0 over all climate regions at different lead times.

3.5 Impact of the weather parameter forecast errors from different NWP

In this section, we analyze how the individual daily weather forecasts (T, RH, Rs and u) from different NWPs (here, only ECMWF and NCEP are shown) influence the daily ET_0 forecasts by replacing one observed weather variable at a time with the forecasts when estimating ET_0.

The results show that solar radiation forecasts have the uppermost impact on ET_0 forecast errors regardless of the forecast system considered (Figure II.12), similarly to what was found by Perera et al. (2014) and Ishak et al. (2010).

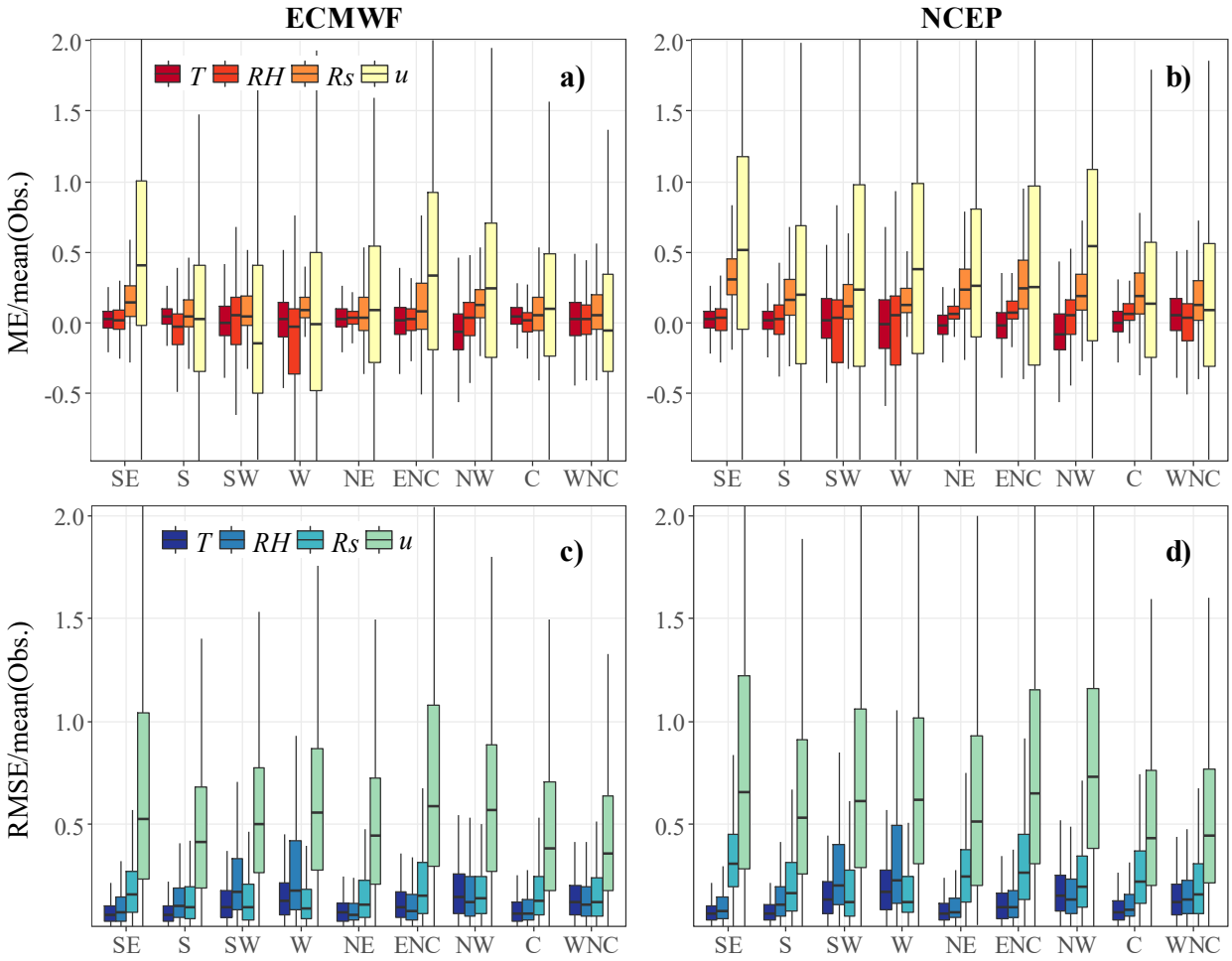

Figure II.12. Box-whisker plots of the scaled ME (a, b) and RMSE (c, d) of ECMWF (a, c) and NCEP (b, d) daily temperature (*T*), relative humidity (*RH*), solar radiation (*Rs*), and wind speed (*u*) forecasts at lead day 3 issued in 2014.

The influence of solar radiation forecasts is dominant even for the regions where the calibrations had minimal effects, such as in the WNC region (Fig. II.13c), suggesting that the impact is also high using calibrated forecasts. Considering, for instance, the large ET_0 errors for the SE region, i.e. the one with the worse performance after calibration, it seems that bad performances in the radiation forecasts are not properly addressed through the post-processing of the ET_0 forecasts. Figures II.12a,b show that both models tend to overestimate the solar radiation forecasts, probably due to model distortions in the representation of surface downwelling

shortwave and longwave radiation (Bodas-Salcedo et al., 2008; Bodas-Salcedo et al., 2012; Mathiesen and Kleissl, 2011; Wild, 2008). Nevertheless, the radiation forecast errors using NCEP are systematically high and are associated with ET_0 forecast errors. Perez et al. (2013) found that the ECMWF model provided significantly better irradiance forecasts than a NCEP-GFS driven mesoscale model for all tested sites in different climate conditions.

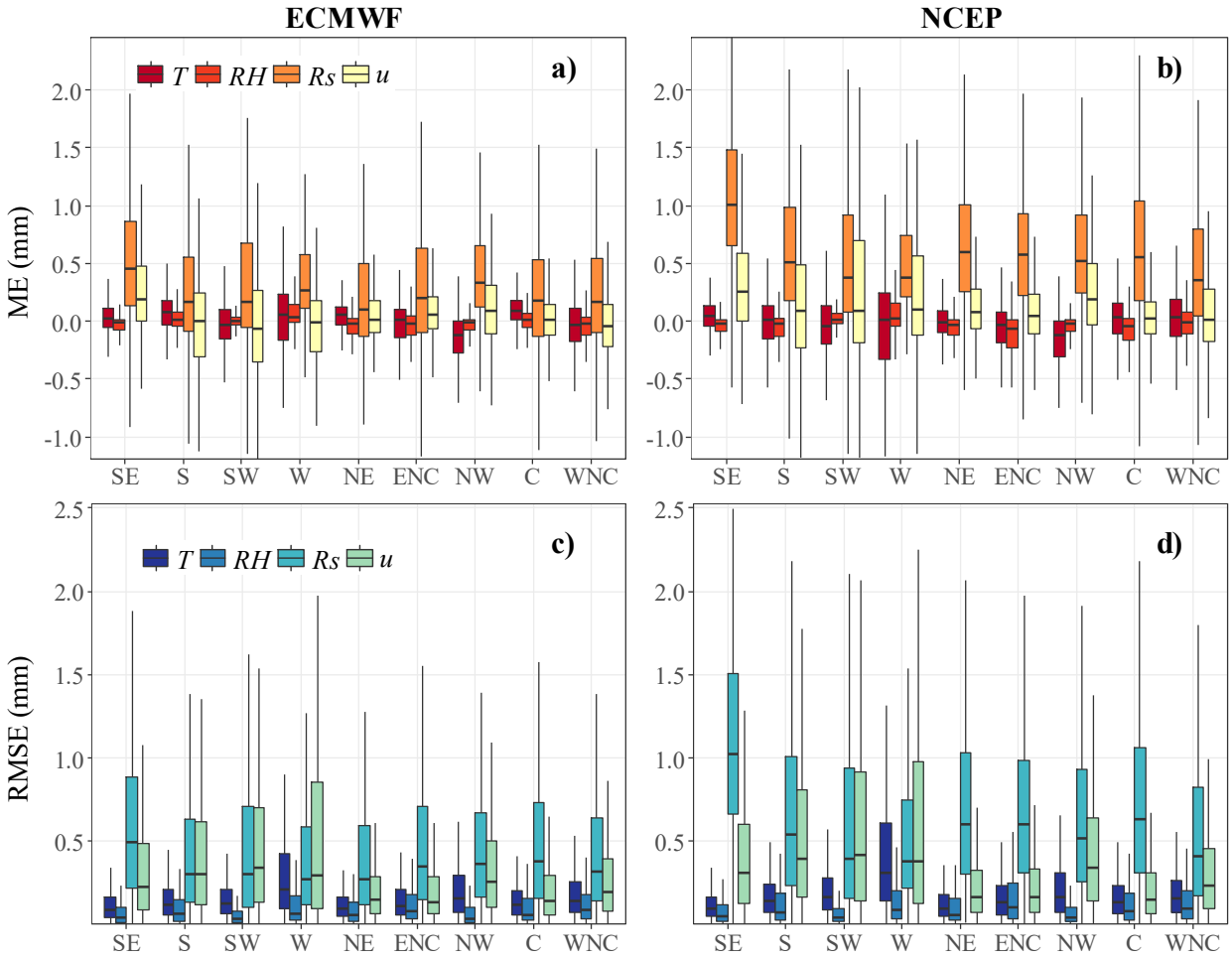

Figure II.13. Box-whisker plots of the scaled ME (a, b) and RMSE (b, c) of ECMWF (a, c) and NCEP (b, d) daily ET_0 forecasts at lead day issued in 2014. The ET_0 forecasts are calculated by replacing one observed variable, either temperature (T), relative humidity (RH), solar radiation (Rs), or wind speed (u) at a time with one corresponding forecast.

Wind speed forecasts show the highest relative errors among weather parameter forecasts, but they commonly have a lower influence on predicted ET_0 values than solar radiation forecasts. Similar results have been reported in other studies (Ishak et al., 2010; Pelosi et al., 2016). These forecasts errors are sensibly higher in the southern regions and especially in the W region, for which the RMSE's of the associated ET_0 forecasts tend to be higher than those relative to the solar

radiation forecasts (Figures II.13c, d). It is worth noting that this study uses one variable at a time replacement method to examine the impact of individual drivers on ET_0 forecast skill. It measures the impact of the individual variables on forecast skill but does not reflect the actual drivers of daily ET_0 variability over CONUS.

Hobbins (2016) provided a decomposition of ET_0 over CONUS variability for each month of the year and showed that during the summer wind speed is the dominant driver over much of the western US and solar radiation is the dominant driver of the southeast. It could explain the fact that the errors in wind speed and solar radiation forecasts have such a detrimental impact on the ET_0 forecasts in the West and Southeast, respectively.

In contrast, temperature forecasts show the lowest bias and highest accuracy among all weather variables and have a limited impact on the ET_0 forecasts, only surpassing that of relative humidity forecasts (Figure II.12). The sensitivity of predicted ET_0 to the errors on T and RH forecasts might be affected by the fact that the biases in minimum and maximum values can compensate each other when evaluating ET_0 with the Penman Monteith equation (Eq. 1). It is worth noting that temperature forecasts contributed most to the ET_0 forecast errors in the W region, suggesting that the unfavorable performance of raw forecasts in this region is determined by the joint effects of large errors in radiation, wind speed and temperature forecasts.

An arising question, considering the contrasting patterns for the ET_0 forecast errors associated with temperature and solar radiation forecasts, is whether simpler ET_0 models, such as the Hargreaves-Samani (Allen et al., 1998; Hargreaves and Samani, 1985), can perform as well as the FAO-56 PM model, by circumventing the errors associated with wind and solar radiation forecasts. Pelosi et al. (2016) found that the Hargreaves-Samani and the Priestley and Taylor models (Priestley and Taylor, 1972) performed well in locations of southern Italy when they were not close to the coastline.

4 CONCLUSIONS

In summary, this study produced and assessed ET_0 forecasts calculated using the leading ECMWF, NCEP, and UKMO model predictions based on FAO-56 PM equation at 101 stations over nine climate regions in CONUS. It examined the probabilistic and deterministic forecasting ability of single-model ensembles as well as two multi-model ensemble systems ECMWF-UKMO and ECMWF-UKMO_NCEP, and also pondered the effects of ensemble member size on a single

model forecast skill. It also identified sources of errors contributing to ET_0 forecasts caused by individual weather forecast variables from different NWP models. This work is helpful for the implementation of reliable operational algorithms for medium-range ET_0 forecasting. It contributes to understand the strengths and weaknesses of three leading forecasting systems when dealing with an agricultural meteorological variable ET_0, which is a complex, nonlinear function of multiple weather parameters.

Major conclusions of this study are highlighted below. First of all, the results revealed that both raw and bias corrected ECMWF forecasts generally had the lowest errors and the highest skill and reliability among all the NWP models being considered, followed by UKMO and NCEP forecasts. The ECMWF-UKMO multi-model ensemble showed consistently better performance than the ECMWF forecasts in experiments where ECMWF and UKMO performed similarly, most comprehending short lead times and/or the W and NW regions. ECMWF-UKMO performed as well as or better than ECMWF-NCEP-UKMO that was found to be unsuitable for implementation. Secondly, the study also showed that even a simple bias correction procedure would remarkably improve the forecast performance, particularly in those regions involving coastal stations or with a complex orography. While the statistical calibration reduced discrepancies between model performances, it did not change the ranking of their performances. Thirdly, the performance of ECMWF forecasts was only slightly influenced by the size of the sampled ensemble when the number of members was equal to or higher than 10, in particular at short lead times. In addition, ECMWF forecasts with only 10 ensemble members provided better performance than the full UKMO forecast ensembles. These results suggested that a statistical calibrated forecast with less members could be more beneficial than using raw forecasts with full ensemble members. Finally, our results suggested that the errors of the radiation forecasts, followed by those of the wind forecasts, had the most detrimental effect on the ET_0 forecasts. Temperature forecasts showed the lowest bias and highest accuracy among all individual weather variables, while contributed slightly more to the ET_0 forecast errors than humidity forecasts. This result suggested that a simpler model with less meteorological input may perform as well as FAO-56 PM model, by avoiding additional errors associated with wind and solar radiation forecasts.

REFERENCES

1. Allen RG., 1996. Assessing integrity of weather data for reference evapotranspiration estimation. J. Irrig. Drain. Eng. ASCE. 122(2):97–106.
2. Allen, R.G., Pereira, L.S., Raes, D., Smith, M., 1998. Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56. FAO, Rome, 300(9): D05109.
3. Allen, R.G., Walter, II.A., Elliott, R., Howell, T., Itenfisu, D., Jensen, M., 2005. The ASCE standardized reference evapotranspiration equation. Rep. 0-7844-0805-X, 59 pp. Available online at http://www.kimberly.uidaho.edu/water/asceewri/ascestzdetmain2005.pdf.
4. Barnston, A.G., Mason, S.J., Goddard, L., DeWitt, D.G., Zebiak, S.E., 2003. Multimodel ensembling in seasonal climate forecasting at IRI. B Am Meteorol Soc, 84(12): 1783-+.
5. Bauer, P., Thorpe, A., Brunet, G., 2015. The quiet revolution of numerical weather prediction. Nature, 525(7567): 47-55.
6. Bodas-Salcedo, A., Ringer, M.A., Jones, A., 2008. Evaluation of the surface radiation budget in the atmospheric component of the Hadley Centre Global Environmental Model (HadGEM1). J Climate, 21(18): 4723-4748.
7. Bodas-Salcedo, A., Williams, K.D., Field, P.R., Lock, A.P., 2012. The Surface Downwelling Solar Radiation Surplus over the Southern Ocean in the Met Office Model: The Role of Midlatitude Cyclone Clouds. J Climate, 25(21): 7467-7486.
8. Buizza, R., 2014. The TIGGE global, medium-range ensembles. European Centre for Medium-Range Weather Forecasts.
9. Buizza, R., Houtekamer, P.L., Pellerin, G., Toth, Z., Zhu, Y., Wei, M., 2005. A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. Mon Weather Rev, 133(5): 1076-1097.
10. Chirico, G.B., Pelosi, A., De Michele, C., Bolognesi, S.F., D'Urso, G., 2018. Forecasting potential evapotranspiration by combining numerical weather predictions and visible and near-infrared satellite images: an application in southern Italy. The Journal of Agricultural Science, pp.1-9. https://doi.org/10.1017/S0021859618000084.

11. Delle Monache, L., Nipen, T., Liu, Y., Roux, G., Stull, R., 2011. Kalman filter and analog schemes to postprocess numerical weather predictions. Mon Weather Rev, 139(11): 3554-3570.
12. Ferro, C.A.T., Richardson, D.S., Weigel, A.P., 2008. On the effect of ensemble size on the discrete and continuous ranked probability scores. Meteorol Appl, 15(1): 19-24.
13. Garcia, M., Raes, D., Allen, R., Herbas, C., 2004. Dynamics of reference evapotranspiration in the Bolivian highlands (Altiplano). Agr Forest Meteorol, 125(1-2): 67-82.
14. Glahn, H.R., Lowry, D.A., 1972. The use of model output statistics (MOS) in objective weather forecasting. J. Appl. Meteorol., 11(8): 1203-1211.
15. Gneiting, T., 2014. Calibration of medium-range weather forecasts. European Centre for Medium-Range Weather Forecasts.
16. Gneiting, T., Balabdaoui, F., Raftery, A.E., 2007. Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69(2): 243-268.
17. Gneiting, T., Raftery, A.E., Westveld III, A.H., Goldman, T., 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Monthly Weather Review, 133(5): 1098-1118.
18. Guo, D., Westra, S., 2017. Modelling Actual, Potential and Reference Crop Evapotranspiration. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.
19. Guo, D.L., Westra, S., Maier, H.R., 2016. An R package for modelling actual, potential and reference evapotranspiration. Environ Modell Softw, 78: 216-224.
20. Hagedorn, R., Buizza, R., Hamill, T.M., Leutbecher, M., Palmer, T.N., 2012. Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts. Q J Roy Meteor Soc, 138(668): 1814-1827.
21. Hagedorn, R., Hamill, T.M., Whitaker, J.S., 2008. Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-meter temperatures. Monthly Weather Review, 136(7), pp.2608-2619.
22. Hamill, T.M., 1997. Reliability diagrams for multicategory probabilistic forecasts. Weather Forecast, 12(4): 736-741.

23. Hamill, T.M., 1999. Hypothesis tests for evaluating numerical precipitation forecasts. Weather Forecast, 14(2): 155-167.
24. Hamill, T.M., 2012. Verification of TIGGE Multimodel and ECMWF Reforecast-Calibrated Probabilistic Precipitation Forecasts over the Contiguous United States. Mon Weather Rev, 140(7): 2232-2252.
25. Hamill, T.M., Bates, G.T., Whitaker, J.S., Murray, D.R., Fiorino, M., Galarneau Jr, T.J., Zhu, Y., Lapenta, W., 2013. NOAA's Second-Generation Global Medium-Range Ensemble Reforecast Dataset. B Am Meteorol Soc, 94(10): 1553-1565.
26. Hamill, T.M., Whitaker, J.S., 2006. Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. Mon Weather Rev, 134(11): 3209-3229.
27. Hamill, T.M., Whitaker, J.S., Mullen, S.L., 2006. Reforecasts - An important dataset for improving weather predictions. B Am Meteorol Soc, 87(1): 33-+.
28. Hargreaves, G.H., Samani, Z.A., 1985. Reference crop evapotranspiration from temperature. Appl Eng Agric, 1(2): 96-99.
29. Hersbach, H., 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather Forecast, 15(5): 559-570.
30. Hobbins, M.T., 2016. The variability of ASCE standardized reference evapotranspiration: A rigorous, CONUS-wide decomposition and attribution. Transactions of the ASABE, 59(2): 561-576.
31. Ishak, A.M., Bray, M., Remesan, R., Han, D., 2010. Estimating reference evapotranspiration using numerical weather modelling. Hydrol Process, 24(24): 3490-3509.
32. Johnson, C., Swinbank, R., 2009. Medium-range multimodel ensemble combination and calibration. Q J Roy Meteor Soc, 135(640): 777-794.
33. Karl, T., Koss, W.J., 1984. Regional and national monthly, seasonal, and annual temperature weighted by area, 1895-1983. National Climatic Data Center.
34. Lawrence, M.G., 2005. The relationship between relative humidity and the dewpoint temperature in moist air - A simple conversion and applications. B Am Meteorol Soc, 86(2): 225-+.
35. Leeper, R.D., Rennie, J., Palecki, M.A., 2015. Observational Perspectives from US Climate Reference Network (USCRN) and Cooperative Observer Program (COOP) Network:

Temperature and Precipitation Comparison. Journal of Atmospheric and Oceanic Technology, 32(4): 703-721.

36. Lewis, C.S., Allen, L.N., 2017. Potential crop evapotranspiration and surface evaporation estimates via a gridded weather forcing dataset. Journal of Hydrology, 546: 450-463.
37. Lewis, C.S., Geli, H.M., Neale, C.M., 2014. Comparison of the NLDAS Weather Forcing Model to Agrometeorological Measurements in the western United States. Journal of Hydrology, 510: 385-392.
38. Mathiesen, P., Kleissl, J., 2011. Evaluation of numerical weather prediction for intra-day solar forecasting in the continental United States. Sol Energy, 85(5): 967-977.
39. Matsueda, M., Endo, H., 2011. Verification of medium-range MJO forecasts with TIGGE. Geophys Res Lett, 38.
40. Menne, M.J., Williams, C.N., Palecki, M.A., 2010. On the reliability of the US surface temperature record. Journal of Geophysical Research: Atmospheres, 115(D11).
41. Mesinger, F., DiMego, G., Kalnay, E., Mitchell, K., Shafran, P.C., Ebisuzaki, W., Jović, D., Woollen, J., Rogers, E., Berbery, E.H., Ek, M.B., 2006. North American regional reanalysis. B Am Meteorol Soc, 87(3): 343-360.
42. Murphy, A.H., 1973. A new vector partition of the probability score. J Appl Meteorol, 12(4): 595-600.
43. NOAA/NESDIS, 2003. United States Climate Reference Network (USCRN). Program Development Plan. CRN series, NOAA-CRN/OSD-2003-0007R0UD0. Available at: https://www1.ncdc.noaa.gov/pub/data/uscrn/documentation/program/X036_d0.pdf.
44. Pelosi, A., Medina, H., Villani, P., D'Urso, G., Chirico, G.B., 2016. Probabilistic forecasting of reference evapotranspiration with a limited area ensemble prediction system. Agr Water Manage, 178: 106-118.
45. Pelosi, A., Medina, H., Van den Bergh, J., Vannitsem, S., Chirico, G.B., 2017. Adaptive Kalman filtering for post-processing ensemble numerical weather predictions. Mon Weather Rev, doi.org/10.1175/MWR-D-17-0084.1
46. Perera, K.C., Western, A.W., Nawarathna, B., George, B., 2014. Forecasting daily reference evapotranspiration for Australia using numerical weather prediction outputs. Agr Forest Meteorol, 194: 50-63.

47. Perez, R., Lorenz, E., Pelland, S., Beauharnois, M., Van Knowe, G., Hemker Jr, K., Heinemann, D., Remund, J., Müller, S.C., Traunmüller, W., Steinmauer, G., 2013. Comparison of numerical weather prediction solar irradiance forecasts in the US, Canada and Europe. Sol Energy, 94: 305-326.
48. Priestley, C.H.B., Taylor, R.J., 1972. On the assessment of surface heat flux and evaporation using large-scale parameters. Monthly weather review, 100(2): 81-92.
49. R Core Team, 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.
50. 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., 2016. The Tigge Project and Its Achievements. B Am Meteorol Soc, 97(1): 49-67.
51. Thielen, J., Bogner, K., Pappenberger, F., Kalas, M., Del Medico, M., De Roo, A., 2009. Monthly-, medium-, and short-range flood warning: testing the limits of predictability. Meteorol. Appl., 16(1): 77-90.
52. Tian, D., Martinez, C.J., 2012a. Comparison of two analog-based downscaling methods for regional reference evapotranspiration forecasts. J. Hydrol., 475: 350-364.
53. Tian, D., Martinez, C.J., 2012b. Forecasting Reference Evapotranspiration Using Retrospective Forecast Analogs in the Southeastern United States. J Hydrometeorol, 13(6): 1874-1892.
54. Tian, D., Martinez, C.J., 2014. The GEFS-based daily reference evapotranspiration (ET_0) forecast and its implication for water management in the southeastern United States. Journal of Hydrometeorology, 15(3): 1152-1165.
55. Tian, D., Martinez, C.J., Graham, W.D., 2014. Seasonal prediction of regional reference evapotranspiration based on climate forecast system version 2. Journal of Hydrometeorology, 15(3), pp.1166-1188.
56. Titley, H., Savage, N., Swinbank, R., Thompson, S., 2008. Comparison between Met Office and ECMWF medium-range ensemble forecast systems. Meteorology Research and Development Technical Report(512).
57. Van Den Dool, H.M., Toth, Z., 1991. Why do forecasts for "near normal" often fail? Weather Forecast, 6(1): 76-85.

58. Wang, K., Augustine, J., Dickinson, R.E., 2012. Critical assessment of surface incident solar radiation observations collected by SURFRAD, USCRN and AmeriFlux networks from 1995 to 2011. Journal of Geophysical Research: Atmospheres, 117(D23).
59. Wild, M., 2008. Short-wave and long-wave surface radiation budgets in GCMs: a review based on the IPCC-AR4/CMIP3 models. Tellus A, 60(5): 932-945.
60. Wilks, D.S., 2006. Comparison of ensemble-MOS methods in the Lorenz'96 setting. Meteorol. Appl., 13(3): 243-256.
61. Wilks, D.S., 2010. Sampling distributions of the Brier score and Brier skill score under serial dependence. Q. J. Roy. Meteor. Soc., 136(653): 2109-2118.
62. Wilks, D.S., 2011. Statistical methods in the atmospheric sciences, 100. Academic press.

CHAPTER III: COMPARISON OF PROBABILISTIC POST-PROCESSING APPROACHES FOR IMPROVING NWP-BASED DAILY AND WEEKLY REFERENCE EVAPOTRANSPIRATION FORECASTS

This chapter has been published in*: Hydrology and Earth System Sciences*, 24(2), pp.1011-1030, 2020.

Abstract: Reference evapotranspiration (ET_0) forecasts play an important role in agricultural, environmental, and water management. This study evaluated probabilistic post-processing approaches, including the nonhomogeneous Gaussian regression (NGR), affine kernel dressing (AKD), and Bayesian model averaging (BMA) techniques, for improving daily and weekly ET_0 forecasting based on single or multiple numerical weather predictions (NWP) from The International Grand Global Ensemble (TIGGE), including the European Centre for Medium-Range Weather Forecasts (ECMWF), the National Centers for Environmental Prediction Global Forecast System (NCEP), and the United Kingdom Meteorological Office forecasts (UKMO). The approaches were examined for the forecasting of summer ET_0 at 101 U.S. Regional Climate Reference Network stations distributed all over the contiguous United States (CONUS). We found that the NGR, the AKD and the BMA methods greatly improved the skill and reliability of the ET_0 forecasts compared to a linear regression bias correction method, due to the considerable adjustments on the spread of ensemble forecasts. The methods were especially effective when applied over the raw NCEP forecasts, followed by the raw UKMO forecasts, because of their low skill compared to that of the raw ECMWF forecasts. The post-processed weekly forecasts had much lower rRMSE (between 8-11%) than the persistence-based weekly forecasts (22%), and the post-processed daily forecasts (13-20%). Compared with the single model ensemble ET_0 forecasts based on ECMWF, multi-model ensemble ET_0 forecasts showed higher skill at short lead times (1 or 2 days) and over the southern and western regions of the United States. The improvement was higher at the daily timescale than at the weekly timescale. The NGR and AKD methods performed the best, but unlike the AKD method, the NGR method can post-process multi-model forecasts and it is easier to interpret than the other methods. In summary, the study demonstrated that the three probabilistic approaches generally outperform conventional procedures based on the simple bias correction of single model forecasts, with the NGR post-processing of the ECMWF and ECMWF-UKMO forecasts providing the most cost-effective ET_0 forecasting.

1 INTRODUCTION

Reference crop evapotranspiration (ET_0) represents the weather-driven component of the water transfer from plants and soils to the atmosphere. It plays a fundamental role in estimating mass and energy balance over the land surface as well as in agronomic, forestry, and water resources management. In particular, ET_0 forecasting is important for aiding water management decision making (such as irrigation scheduling, reservoir operation, etc.) under uncertainty by identifying the range of future plausible water stress and demand (Pelosi et al., 2016; Chirico et al., 2018). While ET_0 forecasts have been mostly focused on the daily timescale (e.g. Perera et al., 2014; Medina et al., 2018), weekly ET_0 forecasts are also important for users. Studies show that both daily and weekly forecasts have an increasing influence on the decision-makers in agriculture (Prokopy et al., 2013; Mase and Prokopy, 2014) and water resource management (Hobbins et al., 2017). For example, irrigation is commonly scheduled considering both a daily and weekly basis, while weekly evapotranspiration forecasts are useful for planning water allocation from reservoirs, especially in cases of shortages. Weekly ET_0 anomalies can also be useful to provide warnings of wild-fires (Castro et al., 2003) and evolving flash drought conditions (Hobbins et al., 2017).

However, ET_0 forecasting is highly uncertain due to the chaotic nature of weather systems. Also, ET_0 estimation requires full sets of meteorological data which are usually not easy to obtain. Due to the improvement of numerical weather predictions (NWPs), studies have been recently emerged to forecast ET_0 using outputs of NWPs over different regions of the world (Silva et al., 2010; Tian and Martinez, 2012 a, 2012b, and 2014; Perera et al., 2014; Pelosi et al., 2016; Chirico et al., 2018; Medina et al., 2018). Operationally, experimental ET_0 forecast products are being developed, such as Forecast Reference EvapoTranspiration (FRET) product (https://digital.weather.gov/), as part of the U.S. National Weather Service (NWS) National Digital Forecast Database (NDFD) (Glahn and Ruth, 2003), and the Australian Bureau of Meteorology's Water and Land website (http://www.bom.gov.au/watl), which provides current and forecasted ET_0 at the continental scale.

The improved performance of NWPs during recent years is largely due to the improvement of physical, statistical representations of the major processes in the models, and the use of ensemble forecasting (Hamill et al., 2013, Bauer et al., 2015). Nevertheless, the NWP forecasts still commonly show systematic inconsistencies with measurements, which are often caused by

inherent errors of NWPs or local land-atmospheric variability which is not well resolved in the models. Post-processing methods, defined as any form of adjustment to the model output to get better predictions (eg., Hagedorn et al., 2012), are highly recommended to attenuate, or even eliminate, those inconsistencies (Wilks, 2006). Until a few years ago, most post-processing applications only considered single-model predictions (i.e., predictions generated by a single NWP model), and addressed errors in the mean of the forecast distribution while ignored those in the forecast variance (Gneiting, 2014). These procedures regularly adopted some form of model output statistics (MOS, Glahn and Lowry, 1972; Klein and Glahn, 1974) methods, focusing on correcting current ensemble forecasts based on the bias in the historical forecasts.

As no forecast is complete without an accurate description of its uncertainty (National Research Council of the National Academies 2006), the dispersion of the forecast ensemble often misrepresents the true density distribution of the forecast uncertainty (Krzysztofowicz 2001; Smith 2001; Hansen 2002). The ensemble forecasts are, for example, commonly under-dispersed (e.g. Buizza et al. 2005; Leutbecher and Palmer, 2008), which make the probabilistic predictions overconfident (Wilks 2011). Therefore, another generation of probabilistic techniques was proposed to also address dispersion errors of the ensembles (Hamill and Colucci 1997; Buizza et al., 2005, Pelosi et al., 2017), in some cases through the manipulation of multi-model weather forecasts.

The nonhomogeneous Gaussian regression (NGR, Gneiting et al., 2005), the Bayesian model averaging, (BMA, Raftery et al., 2005; Fraley et al., 2010), the extended logistic regression (ELR, Wilks et al., 2009; Whan and Schmeits, 2018), the quantile mapping (Verkade et al., 2013) and the family of kernel dressing (Roulston and Smith 2003; Wang and Bishop 2005), such as the affine kernel dressing (AKD, Brocker and Smith 2008), are state of art probabilistic techniques (Gneiting, 2014). However, the ELR has been reported to fall short in using the information contained in the ensemble spread efficiently (Messner et al., 2014), while the quantile mapping method has been found to degrade rather than improve the forecast performance in some circumstances (Madadgar et al., 2014). The NGR, AKD and BMA are sometimes considered as variants of dressing methods (Brocker and Smith 2008), as they produce a continuous forecast probability distribution function (pdf) based on the original ensemble. This property makes them particularly useful for decision making (Gneiting, 2014), compared to the methods that provide post-processed ensembles. Another common advantage is that they perform commonly well with

relatively short training datasets (Geiting et al., 2005; Raftery et al., 2005; Wilks and Hamill, 2007). A limitation of the NGR, compared to the AKD and BMA methods, is that the resulting forecast pdf is invariably Gaussian, while a limitation of the AKD is that it only considers single model ensembles. Instead, the NGR and AKD methods provide more flexible mechanisms for the simultaneous adjustments in the forecast mean and spread-skill (Brocker and Smith, 2008).

Studies suggest that the post-processing of NWP-based ET_0 forecasts are crucial for informing decision making (e.g. Ishak et al., 2010). Medina et al. (2018) compared single and multi-model NWP-based ensemble ET_0 forecasts and the results showed that the performance of the multi-model ensemble ET_0 forecasts is considerably improved through a simple bias-correction post-processing, and that the bias-corrected multi-model ensemble forecasts were in general better than the single model ensemble forecasts. In reality, while most applications for the ET_0 forecasting have involved some form of post-processing, these have been often limited to simple MOS procedures of single-model ensembles (e.g. Silva et al., 2010; Perera et al., 2014). Poor treatment of uncertainty and variability is considered as a main issue affecting users' perceptions and adoptions of weather forecasts (Mase and Prokopy, 2014). The appropriate representation of the second and higher moments of the ET_0 forecast probability density is especially important to predict extreme values, as shown by Williams et al. (2014). Therefore, the use of probabilistic post-processing techniques such as the NGR, the AKD, and BMA, may greatly enhance the overall performance of the ET_0 forecasts compared to the simple MOS procedures.

Only a few studies have considered probabilistic methods for post-processing of ET_0 forecasts. These include the works of Tian and Martinez (2012a, 2012b, and 2014), and more recently Zhao et al (2019). The former authors showed the Analog Forecast (AF) method to be useful for the post-processing ET_0 forecasts based on the Global Forecast System (GFS, Hamill et al., 2006) and the Global Ensemble Forecast System (GEFS, Hamill et al., 2013) reforecasts. Tian and Martinez (2014) found that water deficit forecasts produced with the post-processed ET_0 forecasts had higher accuracy than those produced with climatology. On the other hand, Zhao et al. (2019) improved the skill and the reliability of the Australian BoM model using a Bayesian joint probability (BJP) post-processing approach, which is based on the parametric modeling of the joint probability distribution between forecast ensemble means and observations. However, a main disadvantage of the BJP method compared to the aforementioned state of art probabilistic approaches is that, while they transform the spread of the ensembles, they rely on the mean of

retrospective reforecasts, thus neglecting information about their dispersion. The AF approach has the disadvantage that requires long time series of retrospective forecasts, and may be unsuitable for extreme events forecasting (e.g. Medina et al., 2019). The use of new ET_0 forecasting strategies relying on the postprocessing of single and multi-model ensemble forecasts with the NGR, AKD and the BMA probabilistic techniques provide good opportunities for improving the predictions.

In this paper, we are addressing several scientific questions which have not been adequately studied in previous literature, including, how effective are the state of art probabilistic post-processing methods compared with the traditional MOS bias correction methods for post-processing ET_0 forecasts? Is it worth implementing the probabilistic post-processing for multi-model rather than single-model ensemble forecasting? For the first time, this work aims to evaluate and compare multiple strategies for post-processing both daily and weekly ET_0 forecasts using the NGR, AKD and BMA approaches. The study represents a major step forward with respect to Medina et al. (2018), which evaluated the performance of raw and linear regression bias corrected daily ET_0 forecasts produced with single and multi-model ensemble forecasts. It provides a broad characterization of the performance for different probabilistic post-processing strategies but also diagnoses the causes of high and low performance.

2 METHODS AND DATASETS

2.1 The probabilistic methods

The NGR, AKD and BMA techniques follow a common strategy: they yield a predictive probability density function (PDF) of the post-processed forecasts y given the raw forecasts x and some fitting parameters θ ($p(y|x,\theta)$). The parameters θ are fitted using a training dataset of ensemble forecasts and observations, as in the MOS techniques. Below is a brief description of each technique.

2.1.1 Non-Homogeneous Gaussian Regression

The NGR (Gneiting et al., 2005) produces a Gaussian predictive (PDF) based on the current ensemble (of typically multi-model) forecasts. If x_{ij} denote the j^{th} ($j = 1, \dots, m_i$) ensemble forecast member of model i ($i = 1, \dots, n$), then $p(y|x,\theta) \sim \mathcal{N}(\mu, v)$, where the mean

$$\mu = a + \sum_{i=1}^{n} b_i \bar{x}_i \qquad \text{(III.1)}$$

is a linear combination of the mean ensemble forecasts $\bar{x}_i$ and the variance

$$v = c + dS^2 \quad \text{(III.2)}$$

is a linear function of the ensemble variance S^2. The fitting parameters a, b_i, c and d are determined by minimizing the continuous rank probability score (CRPS) using the training set of forecasts and observations. Notice that parameters a, c, and d are indistinguishable among members; therefore the b_i can be seen as weighting parameters that reflect the better or worse performance of one model compared to the others. The NGR technique is implemented in R (R Core Team) using the packages ensembleMOS (Yuen et al., 2018),

2.1.2. Affine Kernel Dressing

The affine kernel dressing method (Bröcker and Smith, 2008) only considers single model ensemble forecasts. It estimates $p(y|x,\theta)$ using a mixture of normally distributed variables

$$p(y|x,\theta) = \frac{1}{m\sigma}\sum_{j=1}^{m} K\left(\frac{y-z_j}{\sigma}\right) \quad \text{(III.3)}$$

where K represents a standard normal density kernel ($K(\xi) = 1/\sqrt{2\pi}\exp\left(-1/2\xi^2\right)$), centered at z_j, such that

$$z_j = ax_j + r_1 + r_2\bar{x} \quad \text{(III.4)}$$

and,

$$\sigma^2 = h_s^2\left(s_1 + s_2 u(\boldsymbol{z})\right) \quad \text{(III.5)}$$

where h_s is the Silversman's factor (Bröcker and Smith, 2008), $u(\boldsymbol{z})$ is the variance of $\boldsymbol{z}$ and a, r_1, r_2, s_1, s_2 are fitting parameters obtained by minimizing the mean Ignorance score. For clarity we use the same nomenclature for the parameters as in the original study. From Eqs. 4 and 5 we can obtain that the predictive variance v is a function of the ensemble variance S^2 (Brocker and Smith, 2008)

$$v = h_s^2 s_1 + a^2\left(1 + h_s^2 s_2\right)S^2 = c^* + d^* S^2 \quad \text{(III.6)}$$

Here, S^2 represents the variance of the ensemble of exchangeable members. The AKD technique is implemented through the SpecsVerification R package (Siegert, 2017).

2.1.3 Bayesian Model Averaging

The BMA method (Raftery et al. 2005, Fraley et al., 2010) also produces a mixture of normally distributed variables, as the AKD method, but based on multi-model ensemble forecasts.

In this case, the predictive PDF is given by a weighted sum of component PDFs, $g_i(y|x_{i,j};\theta_i\ )$, one per each member:

$$p(y|x,\theta) = \sum_{i=1}^{n}\sum_{j=1}^{m_i} w_i g_i(y|x_{i,j},\theta_i) \quad \text{(III.7)}$$

such that the weights and the parameters are invariable among members of the same model and $\sum_{i=1}^{n} m_i w_i = 1$. In the study the component PDFs are assumed normal as for the affine kernel dressing method. Estimates of w_is and θ_is are produced by maximizing the likelihood function using an Expectation-Maximization algorithm (Casella and Berger, 2002). The BMA technique is implemented through the ensembleBMA R package (Fraley et al., 2016).

2.2 Measurement and forecast datasets

ET_0 observations and forecasts were computed with the FAO-56 PM equation (Allen et al., 1998), from daily meteorological data as inputs, as in Chapter II. They covered the same period, between May and August from 2014 to 2016. The observations used daily measurements of minimum and maximum temperature, minimum and maximum relative humidity, wind speed, and surface incoming solar radiation from 101 U.S. Climate Reference Network (USCRN) weather stations. The USCRN stations are distributed over nine climatologically consistent regions in CONUS (Fig. II. 1). The ET_0 forecasts used daily maximum and minimum temperature, solar radiation, wind speed, and dew point temperature reforecasts of European Centre for Medium-Range Weather Forecasts model (ECMWF) outputs, United Kingdom Meteorological office model (UKMO) outputs, and National Centers for Environmental Prediction model (NCEP) from The International Grand Global Ensemble (TIGGE; Swinbank et al. 2016) database at each of these stations considering a maximum lead time of 7 days. We used the same models as Medina et al. (2018) for comparison purposes, and because they are considered among the most skillful globally (e.g. Hagedorn et al., 2012). The forecasts were interpolated to the same $0.5^o \times 0.5^o$ grid using the TIGGE data portal. The weekly forecasts accounted for the sum of the daily predictions generated at a specific day of each week, and the weekly observations considered the sum of the daily observations over the corresponding forecasting days, such that the weekly observations were independent of each other. In the study, we used the nearest neighbor approach to interpolate the forecasts to the USCRN stations, which does not account for the effects of elevation. While the use of interpolation techniques considering the effects of elevation (e.g. van Osnabrugge et al.,

2019) may correct part of the forecasts errors before the post-processing, it could also affect the multivariate dependence of the weather variables. Hagedorn et al. (2012) showed that the post-processing can not only address the discrepancies related to the model's spatial resolution but also serve as a means of downscaling the forecasts.

2.3 Post-processing schemes

2.3.1 Training and verification periods

The training data for the daily post-processing comprised the pairs of daily forecasts and corresponding observations from 30 days before the forecast initial day, as in Medina et al. (2018). Instead, the training data for the weekly post-processing included all the other pairs of weekly forecasts and observations available for the forecast location, similarly as in the case of a leave one out cross-validation framework. In the study both the daily and weekly forecasts were verified for events over June-August, 2014-2016.

2.3.2 Baseline approaches

Linear regression bias correction (BC) of the ECMWF forecast was used as a baseline approach for measuring the effectiveness of the NGR, the AKD and the BMA methods considering both daily and weekly forecasts. Here, the current forecast bias is estimated as a linear function of the forecasts mean, and the members of the ensemble are shifted accordingly. The function is calibrated using the forecasts mean and the actual biases based on the same training periods as for the other post-processing methods. Persistence is also used as a baseline approach for weekly forecasts, considering its applicability in productive systems. In this case the ET_0 for a current week is estimated as the observed ET_0 during the previous week.

2.3.3 Forecasting Experiments

Table III.1 summarizes the daily and weekly NWP-based ET_0 forecasting experiments based on different post-processing methods and model combinations. The analyses of the daily forecasts put more emphasis on the differences among post-processing methods. They include an examination of the effect of the duration of the training period on the forecast assessments as well as the regression weights from the tested post-processing methods. Whereas, the weekly forecasts

put more emphasis on the differences among the several single and multi-model ET_0 forecasts under baseline and probabilistic post-processing.

2.4 Forecast verification metrics

In this study, we use several metrics to evaluate deterministic and probabilistic forecast performance of the post-processed ET_0 forecasts. For consistency purposes, the metrics of the tested methods were assessed using 50 random samples, i.e., the same number as members in the bias corrected ECMWF forecasts. Deterministic ET_0 forecast was produced by taking the average of the ensemble members. The deterministic forecast performance was assessed using the bias or mean error (ME) and relative ME (rME), the root mean square error (RMSE) and the relative RMSE (rRMSE), and the correlation (ρ), which are common measures of agreement in many studies. The ME and rME were computed as

$$\text{ME} = \frac{1}{n}\sum_{i=1}^{n}\left(\bar{f}_i - o_i\right) \tag{III.8}$$

$$\text{rME} = \frac{\sum_{i=1}^{n}(\bar{f}_i - o_i)}{n\bar{o}} \tag{III.9}$$

where $\bar{f}_i$ represents the average ensemble forecast for the event i (i = 1 … n), o_i is the corresponding observation, and $\bar{o}$ is the mean observed data. The RMSE and the rRMSE were computed as

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\bar{f}_i - o_i\right)^2} \tag{III.10}$$

$$\text{rRMSE} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\bar{f}_i - o_i)^2}}{\bar{o}} \tag{III.11}$$

The correlation was obtained as

$$\rho = \frac{\sum_{i=1}^{n}\left(\bar{f}_i - \bar{\bar{f}}\right)(o_i - \bar{o})}{s_{\bar{f}} s_o} \tag{III.12}$$

where $\bar{\bar{f}}$ is the mean of the average ensemble forecast and $s_{\bar{f}}$ and s_o are the standard deviation of the average forecasts and the observations, respectively.

The probabilistic forecast performance was assessed using range histogram, the spread-skill relationship (see Wilks, 2011) and the forecast coverage as measures of the forecast reliability, the Brier Skill Score (BSS) as a measure of the skill, and the continuous rank probability score

(CRPS), for providing an overall view of the performance (Hersbach, 2000), as it is sensitive to both errors in location and spread simultaneously.

Reliability here refers to the statistical consistency (as in Toth et al. 2003), which is met when the observations are statistically indistinguishable from the forecast ensembles (Wilks, 2011). To obtain the rank histogram, we get the rank of the observation when merged into the ordered ensemble of ET_0 forecasts and then we plot the ranks histogram. The spread-skill relationships are represented as binned-type plots (e.g. Pelosi et al., 2017), accounting for the mean of the ensemble standard deviation deciles (as an indication of the ensemble spread) against the mean RMSE of the forecasts in each decile over the verification period. The plots include the correlation between these two quantities. Calibrated ensembles should show a 1:1 relationship between the standard deviations and the RMSE. If the forecasts are unbiased and the spread is small compared to the RMSE, then the ensembles tend to be under-dispersive. The inverse of the spread provides an indication of sharpness, which is the level of "compactness" of the ensemble (Wilks, 2011).

In addition to the spread skill relationship, we also report the ratio between the observed and nominal coverage (hereinafter referred to as coverage ratio). The coverage of a $(1-\alpha)100\ \%$, $\alpha \in (0,1)$, central prediction interval is the fraction of observations from the verification data set lying between $\alpha/2$ and $1-\alpha/2$ quantiles of the predictive distribution. It is empirically assessed by considering the observations lying between the extreme values of the ensembles. The nominal or theoretical coverage of a calibrated predictive distribution is $(1-\alpha)100\ \%$. A calibrated forecast of m ensemble members provides a nominal coverage of about $(m-1)/(m+1)\ 100\ \%$ central prediction interval (e.g. Beran and Hall, 1993). For example, an ensemble of 50 members provides 96% central prediction interval. The ratio between the observed and nominal coverages provides a quantitative indicator of the quality of the forecast dispersion under unbiasedness: a ratio lower (larger) than 1 suggests that the forecasts tend to be under (over) dispersive.

The BSS is computed as

Table III.1. Evaluated schemes for daily and weekly ET_0 ensemble forecasts with different post-processing methods: BC (simple bias correction), NGR (nonhomogeneous Gaussian regression), AKD (affine kernel dressing), and BMA (Bayesian model averaging), and different model and ensemble schemes: ECMWF, and UKMO ensemble forecasts, as well as ECMWF-UKMO and ECMWF-NCEP-UKMO ensemble forecasts.

	Persistence	BC			NGR			AKD			BMA		
		ECMWF	NCEP	UKMO	ECMWF-UKMO	ECMWF-NCEP-UKMO	ECMWF	ECMWF-UKMO		ECMWF-NCEP-UKMO			
Daily		✓		✓		✓		✓	✓		✓		
Weekly	✓	✓	✓	✓	✓	✓	✓						

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS_{clim}}} \quad \text{(III.13)}$$

where BS is the Brier score of the forecast

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2 \quad \text{(III.14)}$$

p is the forecast probability p of the event, which is estimated based on the ensemble, and o is equal to 1 if the event occurs and 0 otherwise. $\mathrm{BS_{clim}}$ in Eq. 8 represents the Brier Score of the sample climatology, computed as (Wilks, 2010)

$$\mathrm{BS_{clim}} = \bar{o}(1 - \bar{o}) \quad \text{(III.15)}$$

where $\bar{o}$ is the sample climatology computed as the mean of the binary observations o_i in the verification dataset. In this study, we compute the BSS associated with the tercile events of the ET_0 forecasts (upper or 1st, middle or 2nd, and lower or 3rd terciles). Therefore, the sample climatology is equal to $0.3\overline{3}$ and $\mathrm{BS_{clim}} = 0.2\overline{2}$.

The CRPS was computed as

$$\mathrm{CRPS} = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}\left(F_i^f(h) - F_i^o(h)\right)^2 dh \quad \text{(III.16)}$$

where F^f and F^o are the cumulative distribution function of the forecast and the observations, respectively, and h represents the threshold value. $F_i^o(h) = H(h - o_i)$, H representing the Heaviside function, which is 0 for $h < o_i$ and 1 for $h \geq o_i$.

3 RESULTS

3.1 Comparing the NGR, AKD and BMA methods at the daily scale

3.1.1 Deterministic forecast performance

Figure III.1 shows the rME and rRMSE as well as the correlation of the forecasts post-processed using different approaches over the southeast (SE) and northwest (NW) regions. These regions are representative of the Eastern and Western zones, which tended to provide the worse and best rRMSE and correlations, respectively. In general, the probabilistic post-processing methods add no additional skill to the deterministic forecast performance compared to the simple bias correction. While the rRMSE are relatively high, the rME are very low, which indicates that the errors are mostly random. The BMA and the simple linear regression methods provided lower bias than the NGR and AKD methods. Instead, the BMA method provided higher rRMSE and

lower correlations than the other three methods at long lead times. The rRMSE and the correlations tended to be more variable among lead times and regions than among post-processing methods, while for the rME was the opposite. Also, the changes in rRMSE and correlation with lead time tended to be larger over the Eastern regions.

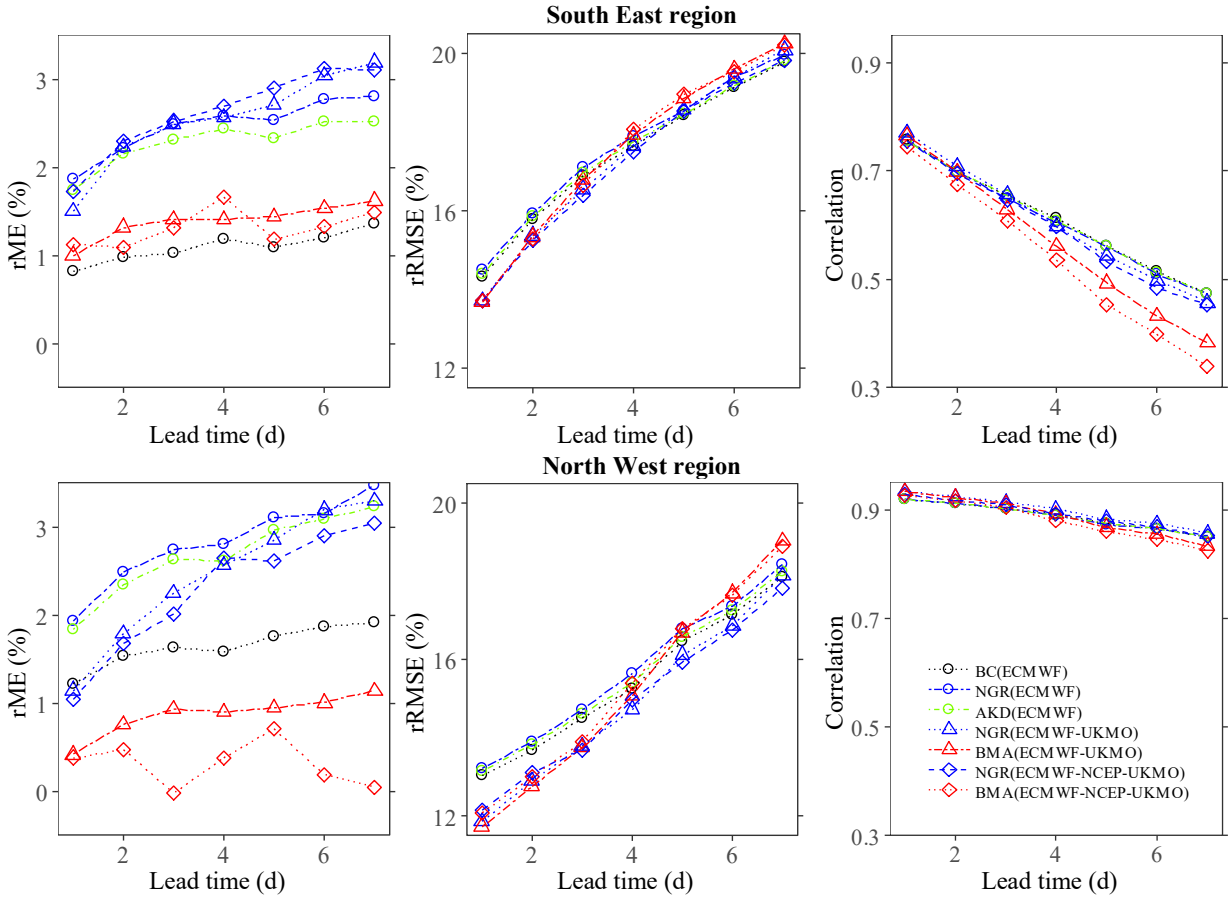

Figure III.1. Relative mean error (rME), relative root mean square error (rRMSE), and correlation considering daily forecasts for different lead times over the SE and NW regions.

3.1.2 Probabilistic forecast performance

Figure III.2 shows the spread skill relationship and the rank histograms using all pairs of forecasts and observations for lead days 1 and 7. The spread-skill relationship shows that the probabilistic post-processing methods considerably improved the reliability of the ET_0 forecasts compared with the linear regression bias correction. The former methods tend to correct evident shortcomings of the ensemble raw forecasts which are unresolved by the simple post-processing, i.e., the considerable under-dispersion at short lead times, and the poor consistency between the ensemble spread and the RMSE at longer lead times. The adjustments had a low cost in terms of

sharpness, judging by the range of ensemble spreads for the different line plots, but seemed slightly insufficient. The correlations between the ensemble standard deviation and the RMSE are fairly low, suggesting a limited predictive ability of the spread (Wilks, 2011). Nonetheless, they were consistently higher for probabilistic post-processing methods, compared to the linear regression method, and at short lead times, compared to the long lead times.

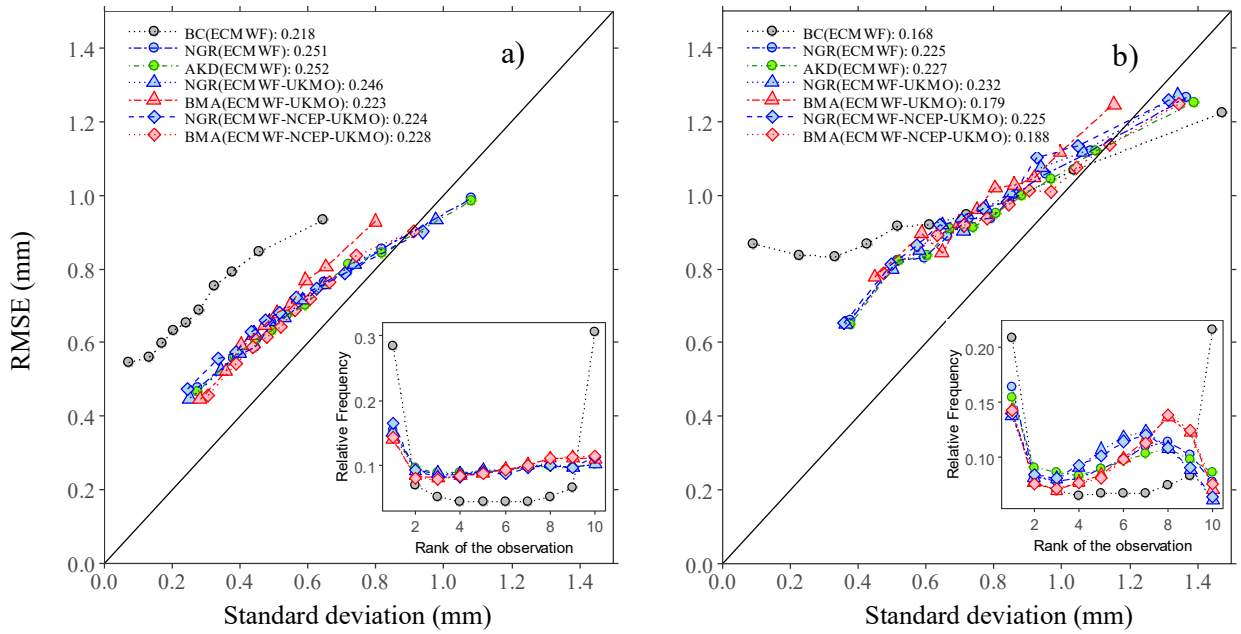

Figure III.2. Binned spread-skill plots accounting for the mean of the ensemble standard deviation deciles against the mean RMSE of the forecasts in each decile over the verification period based on all pairs of forecasts and observations at a) 1-day and b) 7-day lead. The panel in the right and the bottom shows the corresponding rank histograms. The correlation between the standard deviations and the absolute errors is reported after the colon. The solid line represents the 1:1 relationship.

The rank histograms in Figure III.2 show that the probabilistic methods provided better calibration than the linear regression approach both at 1 and 7 days, but the improvements were considerably larger at 1 day. At the short lead time, the three methods slightly over-forecasted ET_0, suggesting that the departures from the predictive mean have a negative skew, but in general they were fairly confident. In this case all the methods provided almost the same result. At the long lead time, there is also an overestimation and then a positive bias, but also a slight U-shaped pattern, associated with some underdispersion for the range of the low and medium observations, which is coherent with the spread skill relationships. These issues are more pronounced using the BMA method and less pronounced using the AKD methods. Scheuerer and Büermann (2014) reported similar issues when post-processing ensemble forecasts of temperatures with the NGR method and

a version of the BMA method. On the other hand, the calibration was affected little by the choice of a single or multi-model strategy for a given post-processing method. Nevertheless, the probabilistic methods provided a coverage ratio close to 100% independently of the lead time (see Table III.2) and the region (not shown). The simple bias correction method instead provided coverage ratios much lower and more variable among regions (see Table III.2) and lead times.

Table III.2. Spatial weighted average values of daily forecast metrics over all climate regions for different methods at lead days 1 and 7. See the caption of Table III.1 for explanations of the methods acronyms. Numbers in bold indicate the best performance for each lead day.

	BC		NGR		AKF		NGR		BMA		NGR		BMA	
	ECMWF		ECMWF		ECMWF		ECMWF-UKMO		ECMWF-UKMO		ECMWF-NCEP-UKMO		ECMWF-NCEP-UKMO	
	1 d	7 d	1 d	7 d	1 d	7 d	1 d	7 d	1 d	7 d	1 d	7 d	1 d	7 d
rME (%)	0.822	1.203	1.695	2.682	1.626	2.419	1.327	2.735	0.632	0.939	1.394	2.778	**0.490**	**0.626**
rRMSE (%)	14.38	**19.64**	14.59	19.88	14.47	19.76	13.68	19.67	13.65	20.15	**13.59**	19.67	13.67	20.28
ME (mm day^{-1})	0.038	0.057	0.080	0.128	0.077	0.115	0.063	0.131	0.029	0.046	0.067	0.134	0.005	0.006
RMSE (mm day^{-1})	0.708	0.950	0.718	0.961	0.716	0.958	0.682	0.965	0.681	0.990	0.681	0.971	0.685	1.002
Correlation	0.832	**0.652**	0.829	0.649	0.830	0.649	**0.843**	0.639	0.841	0.586	0.841	0.635	0.832	0.560
Coverage ratio	64.54	79.40	95.63	95.44	95.93	96.10	94.24	94.73	**96.51**	96.56	93.52	94.57	96.47	**97.24**
CRPS (mm)	0.432	0.555	0.395	0.526	0.394	**0.525**	**0.374**	0.529	0.374	0.547	0.375	0.534	0.377	0.557
BSS_1st	0.442	0.232	0.492	0.279	0.492	**0.282**	**0.525**	0.274	0.519	0.240	0.521	0.271	0.513	0.225
BSS_2nd	0.042	-0.062	0.201	0.101	0.202	**0.101**	**0.224**	0.095	0.214	0.074	0.217	0.089	0.200	0.059
BSS_3nd	0.433	0.300	0.496	**0.359**	0.499	0.358	**0.519**	0.350	0.515	0.305	0.512	0.338	0.494	0.277

The NGR and AFK methods provided better Brier skill score (BSS) than the BC method for the three categories of ET_0 values, with improvements being higher for the middle tercile than for the lower and upper terciles (Figure III.3). The BMA based skill scores tended to decrease with lead time. On west regions (SW, W and NW) and at short lead days the multi-model ensemble forecasts post-processed with the NGR were the most skillful; in the other cases the ECMWF forecasts post-processed with the NGR and the AKD methods tended to be best. The differences of BSS among regions were larger at longer lead times because the skill decreased more sharply over the Eastern regions. This issue is slightly addressed by the NGR and AKD methods based on the ECMWF.

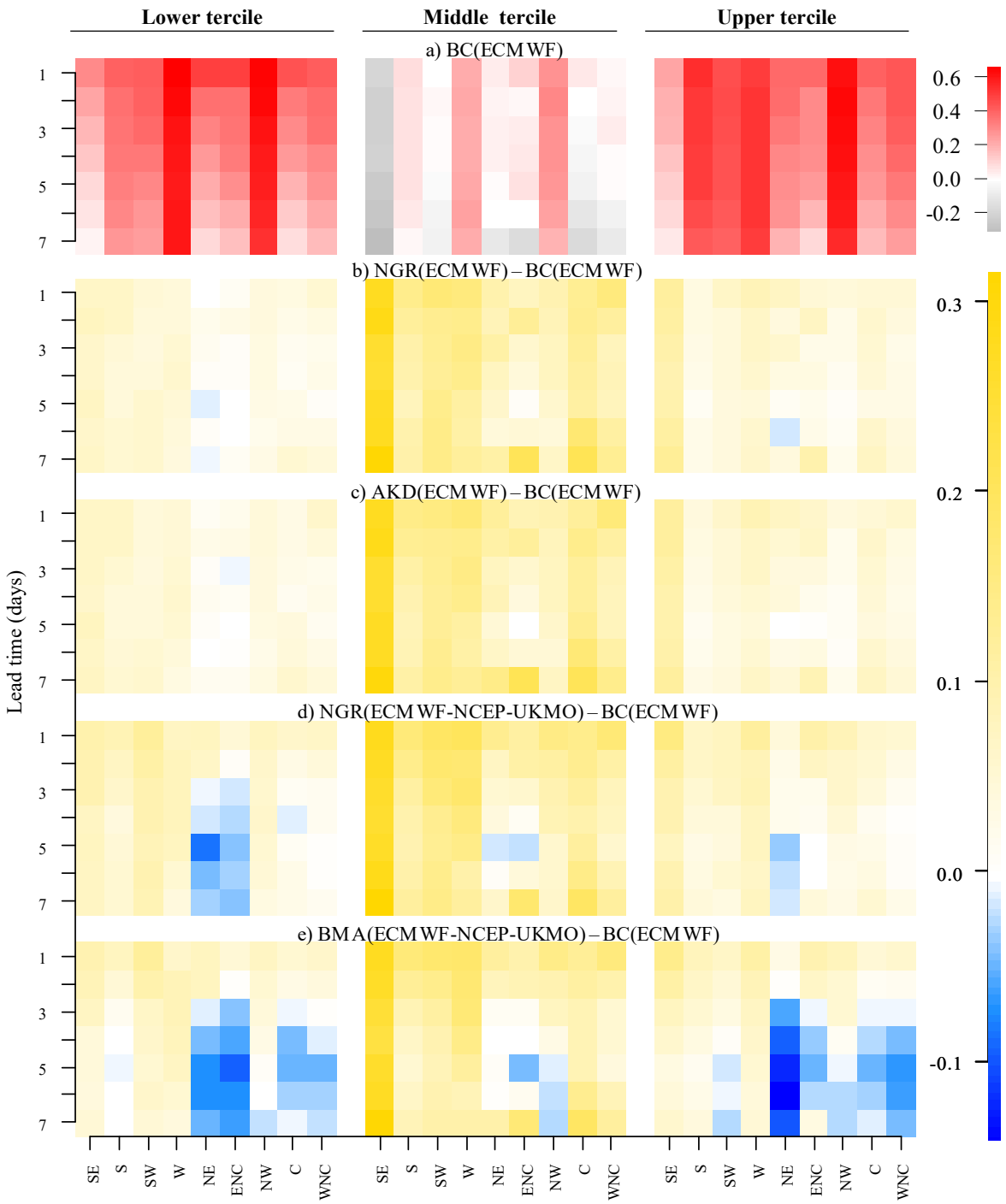

Figure III.3. a) BSS for every region and lead time of the daily ECMWF forecasts post-processed using simple bias correction (used as reference BSS values) and b-e) differences between the BSS of the daily ECMWF forecasts post-processed with the b) NGR and c) AKD methods and the daily ECMWF-NCEP-UKMO forecasts post-processed with the d) NGR and e) BMA methods and the reference BSS.

3.1.3 Summary of average performance for the daily forecast

Table III.2 shows the average performance for the lead days 1 and 7, by weighting the values of each metric according to the number of stations in each region. The ECMWF- UKMO forecasts post-processed with the NGR method were best at short lead times (1-2 days), while the ECMWF forecasts post-processed with the AKD and the NGR methods were the first and second-best at the longer lead times. The BMA method performed well at short lead times but poorly at long times, while the simple bias correction method performed well for deterministic forecasts, but poorly for the probabilistic forecasts. The forecast performance across climate regions is also associated with the choice of the ECMWF ensemble forecasts or the multi-model ensemble forecasts (Table III.3).

Table III.3. Percentage differences (averaged over all lead times) of the ECMWF-UKMO and ECMWF-NCEP-UKMO forecast performance with the ECMWF forecast performance, after post-processing with the non-homogeneous Gaussian regression (NGR) method. See the caption of Table III.1 for explanations of the forecast models' acronyms.

	Western climate regions						Northern climate regions					
	SW		W		NW		NE		ENC		WNC	
	ECMWF-UKMO	ECMWF-NCEP-UKMO	ECMWF-UKMO	ECMWF-NCEP-UKMO	ECMWF-UKMO	ECMWF-NCEP-UKMO	ECMWF-UKMO	ECMWF-NCEP-UKMO	ECMWF-UKMO	ECMWF-NCEP-UKMO	ECMWF-UKMO	ECMWF-NCEP-UKMO
ME	-26.75	-30.83	-9.11	9.42	-13.91	-18.80	-4.27	25.05	-2.15	-1.45	-10.12	0.76
RMSE	-4.68	-4.01	-3.46	-2.51	-3.97	-2.84	1.90	4.33	1.46	2.00	-1.31	-0.92
Correlation	1.76	0.63	0.95	0.71	1.20	0.61	-4.18	-4.60	-3.28	-3.14	-2.31	-2.06
Cov. ratio	-1.39	-2.09	-0.98	-1.19	-1.02	-1.14	-0.84	-1.66	-0.85	-0.99	-0.84	-1.40
CRPS	-4.84	-3.89	-3.42	-1.99	-3.90	-2.81	1.41	4.02	1.58	2.45	-1.00	-0.27
BSS_1st	12.02	7.48	3.22	2.85	3.55	4.24	-12.00	-9.68	-9.64	-9.38	-3.68	-5.18
BSS_2nd	8.99	-6.50	5.79	9.04	4.98	3.96	-112.95	-93.09	-19.09	-13.64	-15.73	-27.95
BSS_3nd	2.30	-1.81	3.58	6.56	4.20	2.37	-9.11	-8.99	-6.42	-10.61	-4.60	-5.84

The single model ECMWF forecasts performed better over northern climate regions than the multi-model ensemble forecasts, while the multi-model did better than any single model forecast over the western regions. The performance over the other regions was more variable among strategies. The performance of the ECMWF- UKMO forecasts was generally better than that of the ECMWF-NCEP- UKMO forecasts (see Table III.3, and Figs. III.1 and III.3). Unlike other performance metrics, the coverage was mostly better for the ECMWF ensemble forecasts than for the multi-model ensemble forecasts. Our CRPS values are comparable with those reported

by Osnabrugge (2019) based on the ECMWF ensemble forecasts of potential evapotranspiration over the Rhine basin, in Europe.

3.1.3 Effect of the length of the training period

The choice of an "optimum" training period is an important issue related to the operational use of post-processing techniques for ET_0 forecasts. Here we compared the performance of different forecasts post-processed with NGR and AKD techniques using 45 and 30 training days. The results suggest that the payoff from using 45 days is practically minimal. Table III.4 shows the percentage differences in the forecasting performance of using 45 and 30 training days for post-processing. While there are generally some minor improvements by using 45 days compared to 30 days, which tend to be higher at longer lead times than shorter times, these improvements usually represent less than 3 percent of original statistics.

Table III.4. Percentage differences (averaged over regions) of forecast performance of using 45 days training period with using 30 days training period for lead days 1 and 7. See the caption of Table III.1 for explanations of the methods acronyms.

	NGR(ECMWF)		AKD(ECMWF)		NGR(ECMWF-UKMO)		NGR(ECMWF-NCEP-UKMO)	
	1 d	7 d	1 d	7 d	1 d	7 d	1 d	7 d
ME	16.57	18.73	21.65	22.86	4.71	10.09	-0.50	7.07
RMSE	-0.70	-2.64	-1.01	-3.12	-0.40	-3.72	-0.05	-4.74
Correlation	-0.16	0.53	-0.14	0.61	-0.10	1.33	-0.47	0.74
Cov. Ratio	1.28	0.95	1.62	1.26	1.70	1.50	1.94	1.34
CRPS (mm)	-0.77	-3.00	-1.22	-3.51	-0.92	-3.89	-0.01	-4.53
BSS_1st	-0.88	2.18	-1.16	2.76	-0.21	5.06	-2.60	6.28
BSS_2nd	-1.26	2.76	-1.28	5.68	3.61	8.96	-2.29	5.56
BSS_3nd	-0.38	-1.59	-0.90	-0.21	-1.34	2.63	-1.63	0.24

The largest percentage difference, accounting for the BSS at the middle tercile, actually represented a negligible gain in absolute terms since they were affected by the close-to-zero range of the variable. The improvements were a bit higher for multi-model ensemble forecasts than for single model forecasts. Notice that, while testing two different periods may be limited to evaluate the methods' sensitivity to the training period, they comprised the range for which methods such as the NGR and BMA have been reported to provide stable results (Gneiting et al., 2005; Raftery et al., 2005).

3.1.4 Weighting coefficients

The weighting coefficients reflect both the performance of the ensemble models and the performance of the post-processing techniques relative to their counterparts. Figure III.4 shows the mean b_i (Eq. 1) weighting coefficients of the NGR technique and w_i (Eq. 7) weighting coefficient of the BMA techniques for each region and lead time for the post-processed ECMWF-NCEP-UKMO, respectively. The coefficients for the NGR and BMA techniques exhibited some common patterns of variability across regions and lead times. Both methods show that the weights of the ECMWF forecasts are at overall the highest, with a clear maximum at medium lead times.

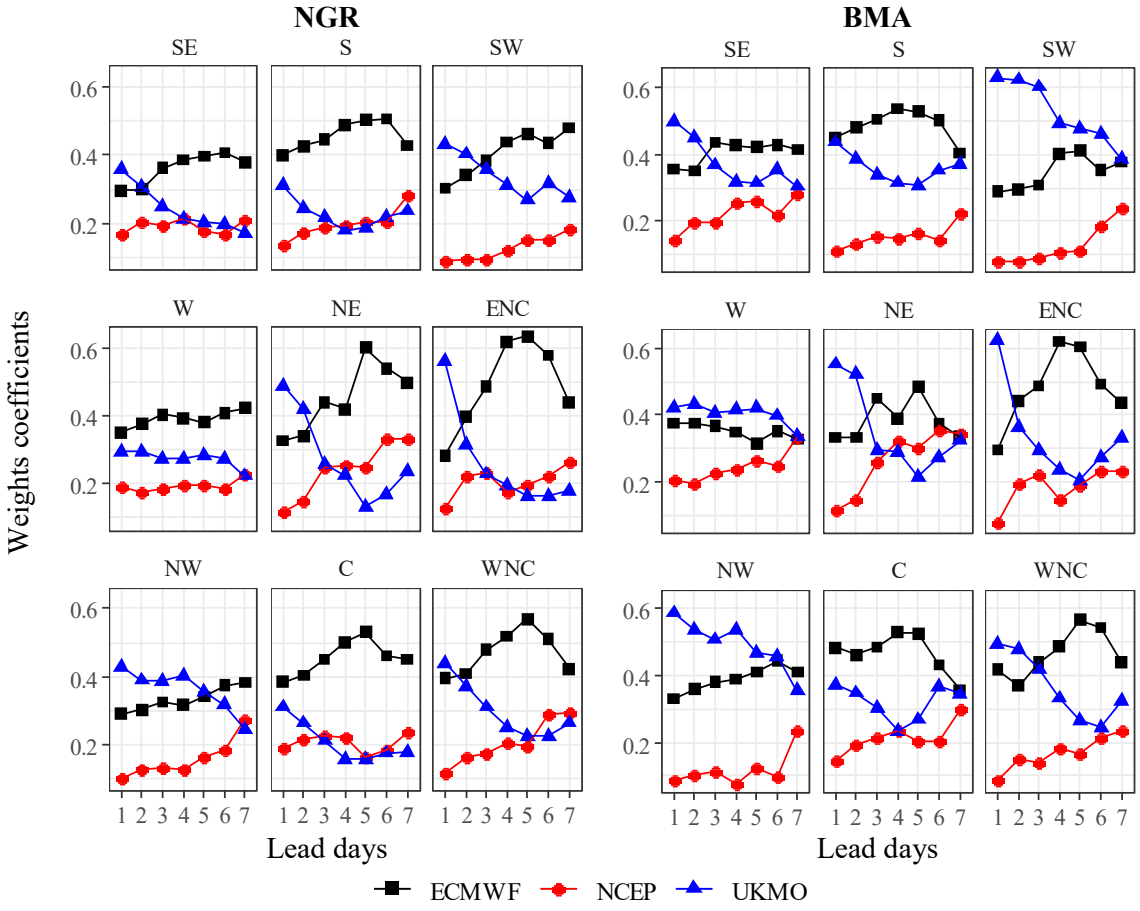

Figure III.4. Regional mean weight coefficient b of the NGR technique (left panel) and the weight coefficient w of the BMA technique (right panel) for the post-processed daily ECMWF-NCEP-UKMO forecasts at different lead days.

The weights of the UKMO model are the highest at 1 and 2 days but sharply decreases with the lead time, while the weights of the NCEP model are in general the lowest, although they consistently increase with lead time, most likely because of the stronger decrease of performance

by the other two models. It explains well the most outstanding features of the performance assessments, concerning the role of each model, and the dependence on regions and lead times. Compared to the NGR method, the BMA method gives the UKMO forecasts a higher relative weight, at the expense of the ECMWF forecast weights. For example, the weighting coefficients of the BMA method over the western regions are consistently higher for the UKMO forecasts than for the ECMWF forecasts. It suggests that the lower performance of the BMA post-processing relative to the NGR and the AKD methods may be related to a misrepresentation of the model weights on the performance. This in turn may be caused by convergence problems during parameter optimization with the expectation-maximization algorithm (Vrugt et al., 2008).

We observed considerable similarities in the distribution of variance coefficients for the NGR method (Eq. 2) and the AKD (Eq. 6) method after post-processing the ECMWF forecasts. The two methods also provide very similar adjustments on the mean forecast because, unlike the BMA method, they independently bias correct the mean and optimize the spread-skill relationship, (Bröcker and Smith, 2008). However, in the experiment the NGR method was about 60 faster than the AKD method. The BMA method was also faster than the AKD method, but still considerably slower than the NGR method. Considering the effectiveness of the NGR method, and its versatility to post-process both single and multi-model ensemble forecasts, we applied this probabilistic technique to weekly ET_0 forecasts based on single model and multi-model ensembles.

3.2 Assessing NGR method for post-processing weekly ET_0 forecasts

3.2.1 Deterministic forecast assessments

As for the daily predictions, the bias, the RMSE and the correlation of the weekly forecasts post-processed with the NGR method and the linear regression methods were similar (Fig. III.5). However, while the RMSE of daily forecasts based on ECMWF model varies between 12 and 20 % of the total ET_0 (Fig. III. 2), the RMSE for any of weekly forecasting strategies commonly varies between 8 and 11%, which is lower than for daily forecasts, making it more useful for operational purposes. The post-processed forecasts showed much lower RMSE and twice higher correlation than the predictions based on persistence, with the weekly predictions based on ECMWF forecasts being generally better, followed by the predictions based on the UKMO forecasts.

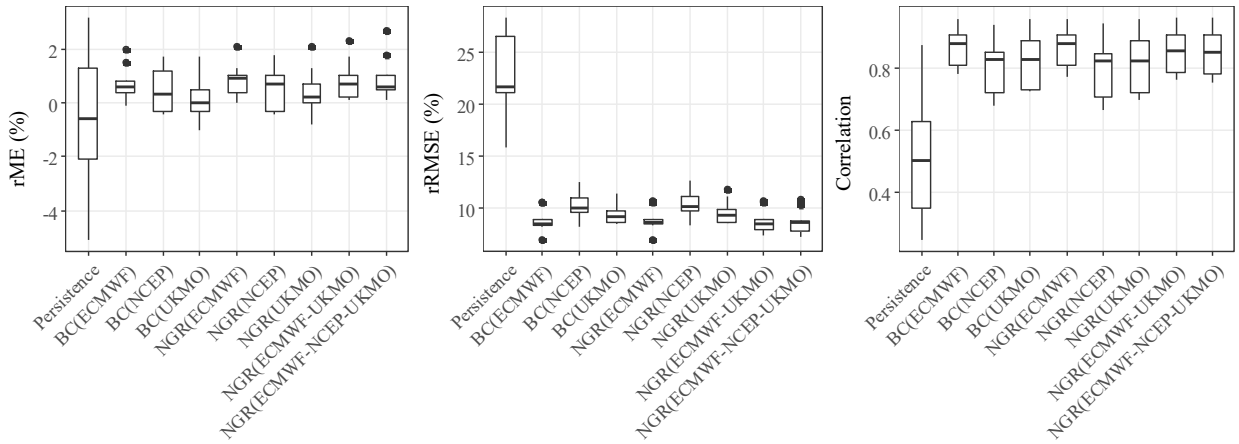

Figure III.5. Whisker plot with the 2.5th, 25th, 50th, 75th and 97.5th percentile of the distribution of the rME, rRMSE and correlation of weekly forecasts across different regions.

3.2.2 Probabilistic forecast assessments

Both the skill and the reliability of the weekly forecasts considerably improved through the NGR post-processing compared with the bias correction post-processing (Table III.5).

Table III.5. Spatial weighted average values of weekly forecast metrics over all climate regions. See the caption of Table III.1 for explanations of the methods acronyms.

	Persistence	BC			NGR				
		ECMWF	NCEP	UKMO	ECMWF	NCEP	UKMO	ECMWF-UKMO	ECMWF-NCEP-UKMO
rME (%)	-0.288	0.683	0.296	**0.097**	0.846	0.496	0.305	0.764	0.814
rRMSE (%)	22.108	8.872	10.453	9.460	8.952	10.571	9.599	8.753	**8.661**
ME (mm week^{-1})	-0.086	0.217	0.077	**0.007**	0.277	0.145	0.080	0.246	0.268
RMSE (mm week^{-1})	7.541	3.059	3.634	3.306	3.086	3.675	3.353	**3.059**	3.064
Correlation	0.530	**0.872**	0.806	0.835	0.870	0.801	0.829	0.863	0.856
Coverage ratio(%)		78.40	48.07	62.92	**99.29**	98.58	98.13	97.74	97.40
CRPS (mm)		1.836	2.406	2.072	1.727	2.071	1.884	**1.708**	1.715
BSS_1st		0.508	0.326	0.448	0.529	0.430	0.501	**0.547**	0.506
BSS_2nd		0.164	-0.147	0.069	0.238	0.150	0.204	**0.255**	0.225
BSS_3nd		0.528	0.371	0.468	0.553	0.461	0.515	**0.558**	0.550

The improvements were different among ET_0 forecast models. In most cases, the better the forecast performance, the lower the improvements are. The adjustments in the coverage ratio and the Brier skill score were about 2.5 and 5 times larger for the UKMO and the NCEP forecasts, respectively than for the ECMWF forecasts. The bias corrected ECMWF forecasts are generally better than both the UKMO and NCEP forecasts post-processed with the NGR method. We found that the post-processing of the NCEP forecasts with methods like the NGR is almost mandatory to

get reasonable probabilistic weekly forecasts of ET_0. For example, the coverage ratio of the bias corrected forecasts on the West region was only 29%, because of the considerable under-dispersion. However, it is notable that, once they were post-processed with the NGR technique, they performed almost comparably to the UKMO forecasts post-processed with the same method, increasing the coverage ratio to 98.4%. Table III.5 also shows that the multi-model ECMWF-UKMO weekly forecasts are commonly the best among all of those post-processed using the NGR method, followed by the ECMWF and the ECMWF-NCEP-UKMO forecasts.

The improvements in the reliability came through substantial adjustments both in the ensemble spread and spread-skill relationship of the raw forecasts (Fig. III.6). The correlations between the standard deviation of the ensembles and the RMSE were more than twice larger through the NGR post-processing than through the linear regression bias correction. The adjustments seemed even slightly more effective than those resulting from the probabilistic post-processing of the daily forecasts (Fig. III. 3), although at the expense of a greater loss of sharpness. The contrasts in the post-processing effectiveness are probably associated with the differences in the training strategies.

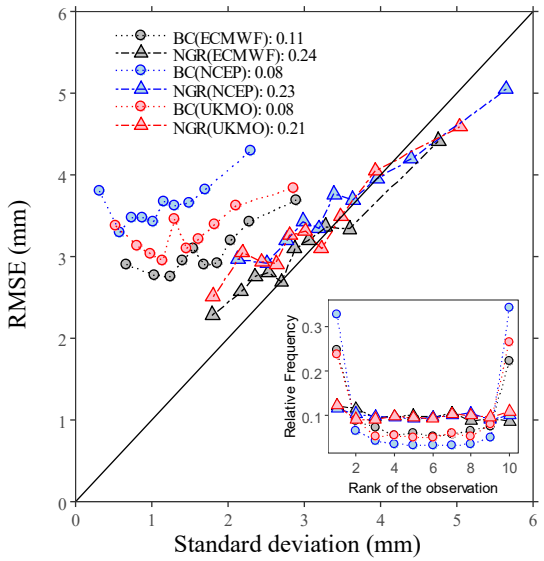

Figure III.6. Binned spread-skill plots for the weekly forecasts accounting for the mean of the ensemble standard deviation deciles against the mean RMSE of the forecasts in each decile over the verification period using all pairs of forecasts and observations. The panel in the right and the bottom shows the corresponding rank histograms. The correlation between the standard deviations and the absolute errors is included in the legend. The solid line represents the 1:1 relationship.

In the case of the probabilistic forecast skill (Fig. III. 7), the improvements were larger for the middle tercile than for the other two terciles, similarly as with daily forecasts. Unlike the bias corrected forecasts, any of the probabilistically post-processed forecasts outperform climatology for practically any tercile and in any region. Maybe, more importantly, the Brier scores for the lower and upper tercile events of the forecasts that have been post-processed with the NGR method is in most cases over 30% better than the scores of climatology. In the coastal regions, from the

South to the Northwest the score is commonly over 50% better, similarly as for the daily forecasts. Finally, the improvements resulting from the use of multi-model ensemble forecasts compared to the single model ensemble forecasts were generally small, except for the Southwest region.

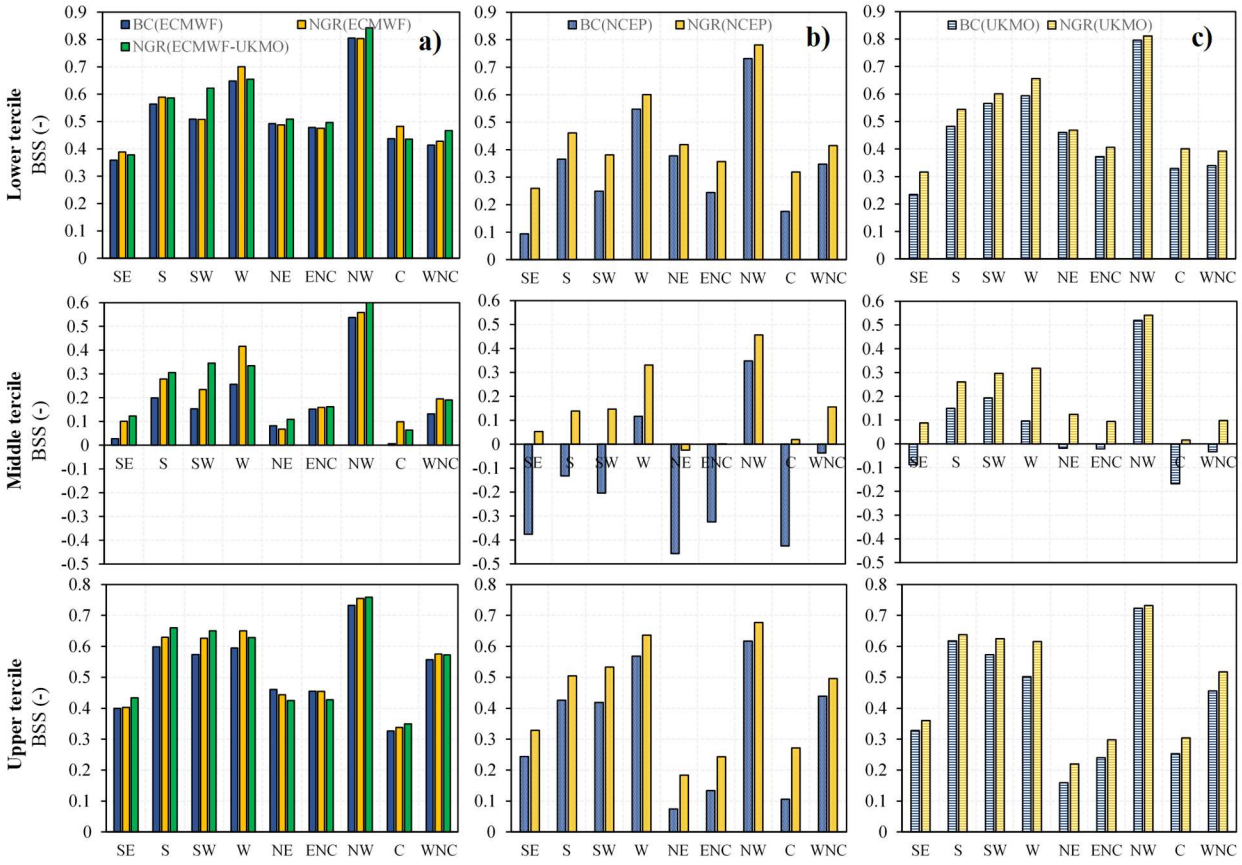

Figure III.7. Comparison between BC and NGR based Brier Skill Scores considering a) ECMWF and ECMWF-UKMO forecasts, b) NCEP, and c) UKMO forecasts across the different climate regions.

4 DISCUSSION

4.1 Effects of probabilistic post-processing on ET_0 forecasting performance

This study showed that NGR, AKD, and BMA post-processing schemes considerably improved the probabilistic forecast performance (coverage ratio, calibration, spread-skill, BSS, CRPS) of the daily and weekly ET_0 forecasts compared with the simple (i.e., using linear regression based on ensemble mean) bias correction method. While sharpness is a wished quality

of any forecast, the daily and weekly bias corrected ET_0 forecasts from NWP are spuriously sharp, which leads to poor consistency between the range of the ET_0 forecasts and the true values, and ultimately undermine the confidence in those forecasts. They also exhibit a poor consistency in that the variance of the ensembles are commonly insensitive to the size of the forecast error. The probabilistic post-processed methods provided much better reliability, with coverage that is close to the nominal value, and at a low cost on sharpness. Therefore, they lead to a much better agreement between the forecasted probability of having an ET_0 event between certain thresholds and the proportions of times that the event occurs (see Gneiting et al., 2005).

In the case of the weekly ET_0 forecasts, the rate of the improvements is considerably smaller for the ECMWF forecasts than for the UKMO, and especially the NCEP forecasts. This seems to be largely due to the better performance of the ECMWF raw forecasts compared to the other forecasting systems. The probabilistic post-processing of the weekly NCEP forecasts seemed practically mandatory to produce reasonable predictions, but once implemented it provided performance assessments almost comparable to those based on the UKMO forecasts. These results have important implications for operational ET_0 forecasts, such as the U.S. national digital forecast database, one of the few operational products of its type, which are based on the NCEP forecasts.

Unlike the probabilistic forecast metrics, the deterministic metrics (ME, RMSE and correlation od the ensemble mean) are low sensitive to the form (deterministic or probabilistic) of post-processing. In particular, the RMSE and correlation seemed more affected by the choice of the single or multi-model ensemble forecast strategy than the choice between the NGR, the AKD or the simple bias correction as a post-processing method. Whereas, RMSE and correlation provided by the BMA method are consistently worse at long lead times. The daily errors under any post-processing were relatively large but mostly random and therefore tend to cancel out at weekly scales. Therefore, while the RMSE varied between 12% and 20% of the daily totals, it represented between 8% and 11% of the weekly totals. The RMSE for weekly ET_0 forecasts were in all cases more than 100% lower than for the persistence-based ET_0 forecasts, and potentially more skillful than the forecasts that exploit the temporal persistence of the ET_0 time-series (e.g. Landeras et al., 2009; Mohan and Arumugam, 2009).

4.2 Comparing the three probabilistic post-processing methods

The NGR and AKD based post-processing methods for the ECMWF forecasts produced comparable results, indicating that the simple Gaussian predictive distribution from the NGR method represents fairly well the uncertainty of the ET_0 predictions. The methods led to a similar distribution of the first two moments of the predictive probability function and similar performance statistics (with the AKD based forecasts being just slightly better). However, the NGR method is more versatile since it can be applied to correct both single model and multi-model ensemble forecasts, while the AKD method can only be applied to correct single model forecasts. The NGR based predictive distribution function is also easier to interpret than the AKD based predictive distribution, which is given by an averaged sum of standard Gaussians.

The BMA method performed slightly less desirable compared to the NGR and AKD presumably due to issues with parameter identifiability. The implemented method uses the Expectation-Maximization (EM) algorithm to produce maximum likelihood estimates of the fitting coefficients, which is susceptible to converge to local minima, especially when dealing with multi-model ensemble forecasts with very different ensemble sizes (Vrugt et al., 2008). Archambeau et al. (2003) demonstrated that, in the presence of outliers or repeated values, this algorithm tends to identify local maximums of the likelihood of the parameters of a Gaussian mixture model. Tian X. et al. (2012) found that adjusted BMA coefficients using both a quasi-Newtonian limited memory algorithm and the Markov Chain Monte Carlo were more accurate than those fitted with the EM algorithm, a procedure that is worth testing in future studies.

4.3 Multi-model ensemble versus single model ensemble forecasts

Daily multi-model ensemble forecasts performed better (in terms of ME, RMSE, correlation, CRPS and BSS) than daily ECMWF forecasts at short lead times (1-2 days) and over the western and southern regions, while the ECMWF forecasts are better over the northeastern regions for longer lead times. For other region/lead time combinations the performance of single and multi-model ensemble forecasts did not differ much. We observed similar patterns for the raw and simple bias corrected forecasts (Medina et al., 2018). Whereas, the weekly multi-model ensemble forecast where consistently better than the weekly single-model forecasts only in the Southwest region, seemingly because the weekly forecasts logically involve both short and long

lead time assessments, and the effectiveness of the multi-models is degraded for long lead times. The observed behavior is associated with the performance of the ECMWF forecasts relative to the UKMO forecasts. While the ECMWF forecasts are in general better than the UKMO and NCEP forecasts, they are much better over the northeastern regions for medium lead times (4-6 days). The UKMO forecasts are in many cases the best at 1 and 2 lead days, but tend to be the worst at the longest times (6-7 days), especially over these regions. The NCEP forecasts had a small contribution compared to the ECMWF and UKMO forecasts at short lead times. These forecasts are comparatively better at longer lead times but still keep a minor role with regard to the ECMWF forecasts.

When considering daily forecasts we adopted a length of the training period of 30 days and showed that by increasing the length to 45 days the improvements were small (commonly lower than three percent). This seems a plausible range for future works and represents an obvious advantage upon methods such as the analog forecast, which provide similar performance (Tian and Martinez 2012 a, b, 2014) but require long training datasets. Gneiting et al. (2005) and Wilson (2007) found that lengths between 30 and 40 days provided good and almost constant performance assessments of sea level pressure forecasts post-processed with the NGR method, and temperature forecasts post-processed with the BMA method, respectively.

4.4. Post-processing the individual inputs versus post-processing ET_0

While in this study we considered the post-processing of ET_0 ensembles produced with raw NWP forecasts, a question is if by post-processing the forcing variables such as temperature, radiation and wind speed first, and then computing the ET_0, we might have better predictions. The NGR method is successful for the post-processing of surface temperatures (e.g. Wilks and Hamill, 2007), whose distribution is fairly Gaussian. For example, Hagedorn (2008) and Hagedorn et al. (2008) showed gains in lead time between two days and four days, with the gains being larger over areas where the raw forecast showed poor skill. Kann et al., (2009) and Kann et al., (2011), used the NGR method for improving short-range ensemble forecasts of 2m-temperature. Recently, Scheuerer and Büermann (2014) provided a generalization of the original approach of Gneiting et al. (2005) that produces spatially calibrated probabilistic temperature forecasts. The wind-speed forecasts have been commonly post-processed with the use of the quantile regression method (e.g. Bremnes 2004; Pinson et al. 2007; Møller et al., 2008). More recently Sloughter et al. (2010)

extended the original BMA method of Raftery et al. (2005) for wind speed, by considering a gamma distribution for modeling the distribution of every member of the ensemble, which considerably improved the CRPS, the absolute errors, and the coverage. Whereas, Vanvyve et al., (2015) and Zhang et al. (2015) used the analog method following the methodology of Delle Monache (2013). The accurate solar radiation forecasting is particularly challenging because it requires a detailed representation of the cloud fields (Verzijlbergh et al., 2015), which is usually not well resolved by the NWP models. Davò et al. (2016) used artificial neural networks (ANN) and the analog method approaches for the post-processing of both wind speed and solar radiation ensemble forecasts, which outperformed a simple bias correction approach. However, the post-processing of meteorological forecasts for producing ET_0 ensemble forecasts may require accounting for the multivariate dependence among those forcing, which is often difficult (e.g. Wilks, 2015). Kang et al (2010) found that post-processing of the streamflow forecasts provided more accurate predictions than post-processing the forcing alone, while Vekade et al (2013) showed that the improvements in precipitation and temperature through the post-processing hardly benefited the streamflow forecasts. Lewis et al., 2014 showed that the performance of the ET_0 forecasts can largely surpass that of the individual input variables. Therefore, it is unclear if we can have any benefit by using the post-processed inputs, instead of the raw forecasts, to construct ET_0 forecasts.

4.5. Future outlook

It is worth noting that, while the ET_0 forecasts are produced for being used in agriculture, they were tested over USCRN stations, which are not representative of agricultural settings. In real applications, the bias between the forecasts with no post-processing and the measurements based on agricultural stations could be higher than the bias resolved in this study. A question that should be addressed in the future studies is to what extent the improvements of the predictive distribution of the ET_0 forecasts can be translated into a more reliable representation of the crop water use in agricultural lands and, ultimately, in water savings and economic gains. Since the ET_0 estimations can have remarkable impacts on the soil moisture estimations (Rodriguez-Iturbe et al., 1999), we envision that new studies relying on the combination of rainfall and ET_0 forecasts post-processed with probabilistic methods will lead to considerable reductions in the uncertainty of soil moisture forecasts. New attempts should also investigate the role of the state of art probabilistic post-

processing techniques on ET_0 forecasts produced from regional numerical weather prediction models, which have had improved spatial resolution and already been used in different meteorological services (e.g. Baldauf et al. 2011; Seity et al. 2011; Hong and Dudhia, 2012; Bentzien and Friederichs, 2012).

5 CONCLUSIONS

This study for the first time evaluated probabilistic methods based on NGR, AKD, and BMA techniques for post-processing daily and weekly ET_0 forecasts derived from single or multi-model ensemble numerical weather predictions. The different ET_0 post-processing methods were compared against the simple linear regression bias correction method using both daily and weekly forecasts, and also against persistence in the case of weekly forecasts. The probabilistic post-processing techniques largely modified the spread of the original ET_0 forecasts, with very favorably impacts on the probabilistic forecast performance. They corrected the notable under-dispersion and the poor consistency between the spread of the ET_0 forecasts and the dimension of the errors, leading to better BSS, reliability (both coverage ratio and spread-skill) and CRPS. The adjustments were crucial on the performance of the weekly NCEP forecasts, followed by the weekly UKMO forecasts, whose bias corrected versions show a clear disadvantage compared with simply post-processed ECMWF forecasts.

The deterministic performance based on the NGR, AKD and BMA methods were comparable to the performance based on the linear regression bias correction for both daily and weekly forecasts, and the skill is about 100% higher than those based on persistence in the case of the weekly forecasts. The rRMSE is between 12 and 20% for the daily totals and 8 and 11% for the weekly totals. The NGR and AKD provided similar estimates of the first and second-order moments of the predictive density distribution; they showed similar effectiveness, but the NGR method has the advantage that can post-process both single and multi-model ensemble forecasts. Both NGR and AKD post-processing methods outperformed the BMA method when considering daily forecasts at long lead times.

The multi-model ensemble forecasting provided benefits at daily scales compared to the ECMWF ensemble forecasting, while the benefits were marginal at weekly scales. The multi-model ensemble forecasting seems a better choice when the UKMO forecasts are comparable or slightly better than the ECMWF forecasts, such as at short (1-2 days) lead times and over the

southern and western regions. Post-processing single model forecast is a better choice than post-processing multi-model ensemble forecast in the circumstances where the ECMWF forecasts perform considerably better than the UKMO and NCEP, such as at mid and long lead times, especially over the northeastern regions. While we considered a length of the training period of 30 days for daily post-processing, the increase of the training period to 45 days only led to minimal improvements. In conclusion, our results suggest that the NGR post-processing of ET_0 forecasts generated from the ECMWF or ECMWF-UKMO predictions is the most plausible strategy among those being evaluated and is recommended for operational implementations because accuracy and reliability requirements for practical applications have not been discussed.

CODE/DATA AVAILABILITY

A repository with the raw and post-processed ET_0 forecasts as well as the R codes used for post-processing is available at http://dx.doi.org/10.17605/OSF.IO/NG6WA.

REFERENCES

1. Allen, R.G., Pereira, L.S., Raes, D., Smith, M., 1998. Crop evapotranspiration-Guidelines for computing crop water requirements-FAO, Irrigation and drainage paper 56, Fao, Rome, 300(9), p.D05109.
2. Archambeau, C., Lee, J.A., Verleysen, M., 2003. On Convergence Problems of the EM Algorithm for Finite Gaussian Mixtures, In ESANN (Vol. 3, pp. 99-106).
3. Baldauf, M., Seifert, A., Förstner, J., Majewski, D., Raschendorfer, M., Reinhardt, T., 2011 Operational convective-scale numerical weather prediction with the COSMO model: Description and sensitivities, Monthly Weather Review, 139(12), pp.3887-3905.
4. Bauer, P., Thorpe, A., Brunet, G., 2015. The quiet revolution of numerical weather prediction, Nature, 525(7567): 47-55.
5. Bentzien, S., Friederichs, P., 2012. Generating and calibrating probabilistic quantitative precipitation forecasts from the high-resolution NWP model COSMO-DE. Weather and Forecasting, 27(4), pp.988-1002.
6. Beran, R., Hall, P., 1993. Interpolated nonparametric prediction intervals and confidence intervals. Journal of the Royal Statistical Society, Series B (Methodological), pp.643-652.

7. Bremnes, J.B., 2004. Probabilistic Wind Power Forecasts Using Local Quantile Regression. Wind Energy, 7, 47–54.
8. Bröcker, J., Smith, L.A., 2008 From ensemble forecasts to predictive distribution functions. Tellus A: Dynamic Meteorology and Oceanography, 60(4), pp.663-678.
9. Buizza, R., Houtekamer, P.L., Pellerin, G., Toth, Z., Zhu, Y., Wei, M., 2005. A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. Monthly Weather Review, 133(5), pp.1076-1097.
10. Casella, G., Berger, R.L., 2002. Statistical inference (Vol. 2). Pacific Grove, CA: Duxbury.
11. Castro, F.X., Tudela, A., Sebastià, M.T., 2003. Modeling moisture content in shrubs to predict fire risk in Catalonia (Spain). Agricultural and Forest Meteorology, 116(1-2), pp.49-59.
12. Chirico, G.B., Pelosi, A., De Michele, C., Bolognesi, S.F., D'Urso, G., 2018. Forecasting potential evapotranspiration by combining numerical weather predictions and visible and near-infrared satellite images: an application in southern Italy. The Journal of Agricultural Science, pp.1-9. https://doi.org/10.1017/S0021859618000084.
13. Davò, F., Alessandrini, S., Sperati, S., Delle Monache, L., Airoldi, D., Vespucci, M.T., 2016. Post-processing techniques and principal component analysis for regional wind power and solar irradiance forecasting. Solar Energy, 134, pp.327-338.
14. Delle Monache, L., Eckel, F.A., Rife, D.L., Nagarajan, B., Searight, K., 2013. Probabilistic weather prediction with an analog ensemble. Monthly Weather Review, 141(10), pp.3498-3516.
15. Fraley, C., Raftery, A.E., Gneiting, T., 2010 Calibrating multi-model forecast ensembles with exchangeable and missing members using Bayesian model averaging. Monthly Weather Review, 138(1), pp.190-202.
16. Fraley, C., Raftery, A.E., Sloughter, J.M., Gneiting T., 2016. EnsembleBMA: Probabilistic Forecasting using Ensembles and Bayesian Model Averaging. R package version 5.1.3. https://CRAN.R-project.org/package=ensembleBMA.
17. Glahn, H.R., Lowry, D.A., 1972. The use of model output statistics (MOS) in objective weather forecasting. J. Appl. Meteorol., 11(8): 1203-1211.
18. Glahn, H.R., Ruth, D.P., 2003. The new digital forecast database of the National Weather Service. Bulletin of the American Meteorological Society, 84(2), pp.195-202.

19. Gneiting, T., Raftery, A.E., Westveld III, A.H., Goldman, T., 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Monthly Weather Review, 133(5), 1098-1118.
20. Gneiting, T., 2014. Calibration of medium-range weather forecasts. European Centre for Medium-Range Weather Forecasts, Technical Memorandum No. 71.
21. Hagedorn, R., Buizza, R., Hamill, T.M., Leutbecher, M., Palmer, T.N., 2012. Comparing TIGGE multi-model forecasts with reforecast-calibrated ECMWF ensemble forecasts. Q J Roy Meteor Soc, 138(668): 1814-1827.
22. Hagedorn, R., Hamill, T.M., Whitaker, J.S., 2008. Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-meter temperatures. Monthly Weather Review, 136, 2608–2619.
23. Hagedorn, R., 2008. Using the ECMWF reforecast data set to calibrate EPS forecasts. ECMWF Newsletter, 117, 8–13.
24. Hamill, T.M., Colucci, S.J., 1997. Verification of Eta–RSM short-range ensemble forecasts. Monthly Weather Review, 125(6), pp.1312-1327.
25. Hamill, T.M., Whitaker, J.S., 2006. Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. Mon Weather Rev, 134(11): 3209-3229.
26. Hamill, T.M., Bates, G.T., Whitaker, J.S., Murray, D.R., Fiorino, M., Galarneau Jr, T.J., Zhu, Y., Lapenta, W., 2013. NOAA's second-generation global medium-range ensemble reforecast dataset. Bulletin of the American Meteorological Society, 94(10), 1553-1565.
27. Hersbach, H., 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather and Forecasting, 15(5), pp.559-570.
28. Hobbins, M., McEvoy, D., Hain, C., 2017. Evapotranspiration, evaporative demand, and drought. Drought and Water Crises: Science, Technology, and Management Issues, pp.259-288.
29. Hong, S.Y., Dudhia, J., 2012 Next-generation numerical weather prediction: Bridging parameterization, explicit clouds, and large eddies. Bulletin of the American Meteorological Society, 93(1), pp.ES6-ES9.
30. Ishak, A.M., Bray, M., Remesan, R., Han, D., 2010. Estimating reference evapotranspiration using numerical weather modelling. Hydrological processes, 24(24), pp.3490-3509.

31. Kang, T.H., Kim, Y.O., Hong, II.P., 2010. Comparison of pre-and post-processors for ensemble streamflow prediction. Atmospheric Science Letters, 11(2), pp.153-159.
32. Kann, A., Haiden, T., Wittmann, C., 2011. Combining 2-m temperature nowcasting and short-range ensemble forecasting. Nonlinear Processes in Geophysics, 18, 903–910.
33. Kann, A., Wittmann, C., Wang, Y., Ma, X., 2009. Calibrating 2-m temperature of limited-area ensemble forecasts using high-resolution analysis. Monthly Weather Review, 137, 3373–3387.
34. Klein, W.H., Glahn, H.R., 1974. Forecasting local weather by means of model output statistics. Bulletin of the American Meteorological Society, 55(10), pp.1217-1227.
35. Landeras, G., Ortiz-Barredo, A., López, J.J., 2009. Forecasting weekly evapotranspiration with ARIMA and artificial neural network models. Journal of irrigation and drainage engineering, 135(3), pp.323-334.
36. Leutbecher, M., Palmer, T.N., 2008. Ensemble forecasting. Journal of Computational Physics, 227(7), pp.3515-3539.
37. Madadgar, S., Moradkhani, H., Garen, D., 2014. Towards improved post-processing of hydrologic forecast ensembles. Hydrological Processes, 28(1), pp.104-122.
38. Mase, A.S., Prokopy, L.S., 2014. Unrealized potential: A review of perceptions and use of weather and climate information in agricultural decision making. Weather, Climate, and Society, 6(1), pp.47-61.
39. Medina, H., Tian, D., Marin, F.R., Chirico, G.B., 2019. Comparing GEFS, ECMWF, and Postprocessing Methods for Ensemble Precipitation Forecasts over Brazil. Journal of Hydrometeorology, 20(4), pp.773-790.
40. Medina, H., Tian, D., Srivastava, P., Pelosi, A., Chirico, G.B., 2018. Medium-range reference evapotranspiration forecasts for the contiguous United States based on multi-model numerical weather predictions. Journal of Hydrology, 562, pp.502-517.
41. Messner, J.W., Mayr, G.J., Zeileis, A., Wilks, D.S., 2014. Heteroscedastic Extended Logistic Regression for Postprocessing of Ensemble Guidance. Mon. Wea. Rev., 142, 448–456, https://doi.org/10.1175/MWR-D-13-00271.1.
42. Mohan, S., Arumugam, N., 1995. Forecasting weekly reference crop evapotranspiration series. Hydrological Sciences Journal, 40(6), pp.689-702.
43. Møller, J.K., Nielsen, H.A., Madsen, H., 2008. Time-Adaptive Quantile Regression, Computational Statistics & Data Analysis, 52, 1292–1303.

44. National Research Council of the National Academies, 2006. Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts. The National Academies Press, 124 pp.
45. Osnabrugge, B.V., Uijlenhoet, R., Weerts, A., 2019. Contribution of potential evaporation forecasts to 10-day streamflow forecast skill for the Rhine River. Hydrology and Earth System Sciences, 23(3), pp.1453-1467.
46. Pelosi, A., Medina, H., Van den Bergh, J., Vannitsem, S., Chirico, G.B., 2017. Adaptive Kalman filtering for post-processing ensemble numerical weather predictions. Mon Weather Rev, doi.org/10.1175/MWR-D-17-0084.
47. Pelosi, A., Medina, H., Villani, P., D'Urso, G., Chirico, G.B., 2016. Probabilistic forecasting of reference evapotranspiration with a limited area ensemble prediction system. Agricultural water management, 178, pp.106-118.
48. Perera, K.C., Western, A.W., Nawarathna, B., George, B., 2014. Forecasting daily reference evapotranspiration for Australia using numerical weather prediction outputs. Agr. Forest Meteorol., 194: 50-63.
49. Pinson, P., Madsen, H., 2009. Ensemble-Based Probabilistic Forecasting at Horns Rev. Wind Energy, 12, 137–155.
50. Prokopy, L.S., Haigh, T., Mase, A.S., Angel, J., Hart, C., Knutson, C., Lemos, M.C., Lo, Y.J., McGuire, J., Morton, L.W., Perron, J., 2013. Agricultural advisors: a receptive audience for weather and climate information? Weather, Climate, and Society, 5(2), pp.162-167.
51. R Core Team, 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/.
52. Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian model averaging to calibrate forecast ensembles, Monthly Weather Review, 133(5), pp.1155-1174.
53. Rodriguez-Iturbe, II., Porporato, A., Ridolfi, L., Isham, V., Coxi, D.R., 1999. Probabilistic modelling of water balance at a point: the role of climate, soil and vegetation. Proceedings of the Royal Society of London, Series A: Mathematical, Physical and Engineering Sciences, 455(1990), pp.3789-3805.
54. Roulston, M.S., Smith, L.A., 2003. Combining dynamical and statistical ensembles. Tellus A: Dynamic Meteorology and Oceanography, 55(1), pp.16-30.

55. Scheuerer, M., Büermann, L., 2014. Spatially adaptive post-processing of ensemble forecasts for temperature. Journal of the Royal Statistical Society: Series C (Applied Statistics), 63(3), pp.405-422.
56. Seity, Y., Brousseau, P., Malardel, S., Hello, G., Bénard, P., Bouttier, F., Lac, C., Masson, V., 2011. The AROME-France convective-scale operational model. Monthly Weather Review, 139(3), pp.976-991.
57. Siegert, S., 2017. SpecsVerification: Forecast Verification Routines for Ensemble Forecasts of Weather and Climate. R package version 0.5-2. https://CRAN.R-project.org/package=SpecsVerification.
58. Silva, D., Meza, F.J., Varas, E., 2010. Estimating reference evapotranspiration (ET_0) using numerical weather forecast data in central Chile. Journal of hydrology, 382(1-4), pp.64-71.
59. Sloughter, J.M., Gneiting, T., Raftery, A.E., 2010. Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. Journal of the american statistical association, 105(489), pp.25-35.
60. 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., 2016. The Tigge Project and Its Achievements. B. Am. Meteorol. Soc., 97(1): 49-67.
61. Tian, D., Martinez, C.J., 2012a. Comparison of two analog-based downscaling methods for regional reference evapotranspiration forecasts. J. Hydrol., 475: 350-364.
62. Tian, D., Martinez, C.J., 2012b. Forecasting Reference Evapotranspiration Using Retrospective Forecast Analogs in the Southeastern United States. J. Hydrometeorol., 13(6): 1874-1892.
63. Tian, D., Martinez, C.J., 2014. The GEFS-based daily reference evapotranspiration (ET_0) forecast and its implication for water management in the southeastern United States. J. Hydrometeorol., 15(3): 1152-1165.
64. Tian, X., Xie, Z., Wang, A., Yang, X., 2012. A new approach for Bayesian model averaging. Science China Earth Sciences, 55(8), 1336-1344.
65. Toth, Z., Talagrand, O., Candille, G., Zhu, Y., 2003. Probability and ensemble forecasts, Forecast Verification: A Practitioner's Guide in Atmospheric Science, pp.137-163.
66. Vanvyve, E., Delle Monache, L., Monaghan, A.J., Pinto, J.O., 2015. Wind resource estimates with an analog ensemble approach. Renewable Energy, 74, pp.761-773.

67. Verkade, J.S., Brown, J.D., Reggiani, P., Weerts, A.H., 2013. Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales. Journal of Hydrology, Volume 501,2013, Pages 73-91,http://dx.doi.org/10.1016/j.jhydrol.2013.07.039.
68. Verkade, J.S., Brown, J.D., Reggiani, P., Weerts, A.H., 2013. Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales. Journal of Hydrology, 501, pp.73-91.
69. Verzijlbergh, R.A., Heijnen, P.W., de Roode, S.R., Los, A., Jonker, H.J., 2015. Improved model output statistics of numerical weather prediction based irradiance forecasts for solar power applications. Solar Energy, 118, pp.634-645.
70. Vrugt, J.A., Diks, C.G., Clark, M.P., 2008. Ensemble Bayesian model averaging using Markov chain Monte Carlo sampling. Environmental fluid mechanics, 8(5-6), pp.579-595.
71. Wang, X., Bishop, C.H., 2005. Improvement of ensemble reliability with a new dressing kernel. Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography, 131(607), pp.965-986.
72. Whan, K., Schmeits, M, 2018. Comparing Area Probability Forecasts of (Extreme) Local Precipitation Using Parametric and Machine Learning Statistical Postprocessing Methods. Mon. Wea. Rev., 146, 3651–3673, https://doi.org/10.1175/MWR-D-17-0290.1.
73. Wilks, D.S., Hamill, T.M., 2007 Comparison of ensemble-MOS methods using GFS reforecasts. Monthly Weather Review, 135(6), pp.2379-2390.
74. Wilks, D.S., 2006 Comparison of ensemble-MOS methods in the Lorenz'96 setting. Meteorological Applications, 13(3), pp.243-256.
75. Wilks, D.S., 2009. Extending logistic regression to provide full probability distribution MOS forecasts. Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling, 16(3), pp.361-368.
76. Wilks, D.S., 2015. Multivariate ensemble Model Output Statistics using empirical copulas. Quarterly Journal of the Royal Meteorological Society, 141(688), pp.945-952.
77. Wilks, D.S., 2010. Sampling distributions of the Brier score and Brier skill score under serial dependence. Q. J. Roy. Meteor. Soc., 136(653): 2109-2118.

78. Williams, R.M., Ferro, C.A.T., Kwasniok, F., 2014. A comparison of ensemble post-processing methods for extreme events. Quarterly Journal of the Royal Meteorological Society, 140(680), pp.1112-1120.
79. Wilson, L.J., Beauregard, S., Raftery, A.E., Verret, R., 2007. Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian model averaging. Monthly Weather Review, 135(4), pp.1364-1385.
80. Yuen, R., Baran, S., Fraley, C., Gneiting, T., Lerch, S., Scheuerer, M., Thorarinsdottir, T., 2018. ensembleMOS: Ensemble Model Output Statistics. R package version 0.8.2. https://CRAN.R-project.org/package=ensembleMOS.
81. Zhang, J., Draxl, C., Hopson, T., Delle Monache, L., Vanvyve, E., Hodge, B.M., 2015. Comparison of numerical weather prediction based deterministic and probabilistic wind resource assessment methods. Applied Energy, 156, pp.528-541.
82. Zhao, T., Wang, Q.J., Schepen, A., 2019. A Bayesian modelling approach to forecasting short-term reference crop evapotranspiration from GCM outputs. Agricultural and Forest Meteorology, 269, pp.88-101.

CHAPTER IV: COMPARING NCEP, ECMWF, AND POST-PROCESSING METHODS FOR ENSEMBLE PRECIPITATION FORECASTS OVER BRAZIL

This chapter has been published in: *Journal of Hydrometeorology*, 20(4), pp.773-790, 2019.

Abstract: This study compares precipitation ensemble forecasting of medium-range National Centers for Environmental Prediction (NCEP), European Center Medium Range Weather Forecasts (ECMWF), and NCEP post-processed with six analog-based methods and a logistic regression method over six biomes in Brazil. The numerical weather prediction (NWP) forecasts were evaluated against the Physical Science Division South America Daily Gridded Precipitation dataset using both deterministic and probabilistic forecasting evaluation metrics. The results show that the ensemble precipitation forecasts performed commonly well in the East and poorly in the Northwest of Brazil, independent of the models and the post-processing methods. While the raw ECMWF forecasts performed better than the raw NCEP forecasts, analog-based NCEP forecasts were more skillful and reliable than both raw ECMWF and NCEP forecasts. The choice of a specific post-processing strategy had less impact on the performance than the post-processing itself. Nonetheless, forecasts produced with different analog-based post-processing strategies were significantly different and were more skillful and as reliable and sharp as forecasts produced with the logistic regression method. The approach considering the logarithm of current and past reforecasts as the measure of closeness between analogs was identified as the best strategy. The results also indicate that the post-processing using analog methods with long-term reforecast archive improved raw NCEP precipitation forecasting skill, more than using logistic regression with short-term reforecast archive. In particular, the post-processing dramatically improves the NCEP precipitation forecasts when the forecasting skill is low or below zero.

1 INTRODUCTION

Precipitation is a major source of water resources and a determinant in the functioning of agriculture, forest and freshwater ecosystems. Accurate precipitation forecasting is one of the most sensible aspects of weather prediction for society. It strongly affects daily decisions in different sectors, such as public health, water resources, energy production, agriculture, and environmental protection. Numerical weather prediction models (NWP) is the state-of-art technology for forecasting medium-range precipitation at daily or sub-daily time step over the globe. The improvements in resolution, parameterization, and physical representation of the main processes and phenomena, has prompted the use of medium-range NWP forecasts in many weather-dependent activities (Hamill et al., 2013). The skill of medium-range forecasts has increased by about one day per decade, meaning that today's 6-day forecast is as accurate as the 5-day forecast ten years ago (Bauer et al., 2015).

NWP has global applicability (Bauer et al., 2015) and potential for improving regional precipitation, run-off, and water storage forecasting over the globe (e.g. Hamill et al., 2012, Su et al., 2014; Wetterhall et al., 2010; Cloke and Pappenberger, 2009). However, few studies have focused on assessing the NWP precipitation predictability associated with large and intense mesoscale convective systems (Bechtold et al., 2012), such as tropical rainfall. In reality, atmospheric convection plays a key role in regulating the climate in the tropics (Bony et al., 2015), and is one of the most challenging processes to parametrize in weather and climate models (Bauer, 2015). The challenges have been greater over continental areas from the Southern Hemisphere where the abundant vegetation and the sparse observations for evaluation and data assimilation have limited the models' accuracy. Recent progress in forecasting tropical convection (Bechtold et al., 2014; Subramanian et al., 2017) and the increasing quantity and quality of global information encourage the use of NWP for tropical precipitation forecasting. It is, therefore, necessary to conduct comprehensive assessments of the NWP's ability to forecast heavy and highly variable rainfall regimes in tropical and near tropical regions dominated by large mesoscale convective systems (Mohr and Zipser, 1996).

The National Centers for Environmental Prediction (NCEP) Global Ensemble Forecast System (GEFS), and the European Centre for Medium-Range Weather Forecasts the ensemble prediction system (ECMWF) are two leading NWPs for medium-range weather forecasting at the

global scale. In particular, the ECMWF global ensemble forecasts have consistently been the most skillful among those produced at national weather forecast centers during the last decades (e.g., Buizza et al. 2005, Hagedorn et al., 2012). An advantage of the NCEP model is that it archives retrospective forecast (reforecast) data sets for long past periods at no cost, which are useful for statistically post-processing to correct weather forecasts against observed data, thus reducing the uncertainty and improving forecast performance (Hamill et al. 2006; Hagedorn 2008). Statistical post-processing methods often succeed to reduce forecast errors and account for local meteorological conditions that are not resolved at the spatial scale of the NWP model grid (e.g. Glahn and Lowry, 1972; Gneiting, 2014; Pelosi et al., 2017). However, it is still not well understood the relative performance of NCEP, ECMWF, and statistically post-processed precipitation forecasts in the tropical and near tropical regions dominated by large and intense mesoscale convective systems.

The analog post-processing method is an efficient approach to improve probabilistic precipitation forecasts (Voisin et al., 2010, Daoud et al., 2016) and in general several other hydrometeorological forecasts (Tian and Martinez, 2012, 2014). In this method, the current forecast from a fixed NWP is compared against the past forecasts of the same NWP at a similar time of the year within a limited region, and an ensemble is formed considering the observations on the dates of the closest matches (Hamill et al., 2006). Studies have explored different strategies for implementing analog methods with NCEP reforecast, such as testing different similarity criteria (Hamill and Whitaker, 2006), and multivariate (Hamill and Whitaker, 2006; Delle Monache et al. 2011; 2013) versus univariate similarity metrics, and evaluating different sizes of the search region (Hamill and Whitaker 2006; Hamill et al., 2015; Tian and Martínez 2012, 2014) and number of ensemble members (Hamill et al., 2015). Nevertheless, guidelines regarding the optimal implementing strategies to efficiently post-processing tropical convective precipitations are still lacking.

A disadvantage of the analog approaches is that it needs long-term reforecasts for finding the closest matching analogs. When the forecasted precipitation is a large, rare event, it becomes a challenge to find a sufficient number of analogs if the reforecast archive is not sufficiently long enough (Hamill, 2015). There are alternative approaches that are less reliant on the size of the training data. The Logistic Regression method is one of these methods and has been found suitable for dealing with medium-range precipitation forecasts in several regions (Wilks 2006; Wilks and

Hamill, 2007). Few previous studies have compared the relative performance of analog techniques and logistic regression techniques for post-processing NCEP precipitation forecasts. For selecting optimal post-processing methods, it would be informative to compare the performance of analog methods, which requires long-term reforecast archives, with logistic regression, which only needs a small set of training data.

Given the research gaps we have identified, this study was aimed to: 1) document the performance of the NCEP and ECMWF daily precipitation ensemble forecasts using Brazil as a study case, 2) evaluate the NCEP-based precipitation forecasts post-processed using analog methods with different strategies, and 3) compare the performance of Analog-based methods with the Logistic Regression method.

Brazil covers a large area and is considerably affected by large and intense mesoscale convective systems within which severe weather events develop (Mohr and Zipser, 1996). The complexity of the spatial and temporal variability of rainfall patterns over Brazil may provide a unique setting for assessing progresses of global scale NWPs and post-processing techniques for rainfall prediction.

2 MATERIALS AND METHODS

2.1 Study region

Brazil is of a mega-diverse and the world's fifth-most populous country. It is the second country with the largest forest area in the world (FAO, 2015), a country with high risks of vector-borne transmission diseases (WHO, 2014), one of the top hydropower potential countries (Zhou et al. 2015), and one of the world's main producers of food and biofuels (Ferreira et al., 2012). It ranks first in sugarcane, coffee, or oranges productions and sixth in the world cereal production (FAO 2014). Given the significant impact of precipitation in those sectors, forecasting medium-range daily precipitation for Brazil will have great implications for its agriculture, natural resources, hydropower generation, and public health management. The study focused on the six major natural biomes of Brazil: Amazon, Caatinga, Cerrado, Atlantic Forest, Pampa and Pantana, representing climatologically consistent regions (Figure IV.1). A brief description of each biome is provided as follows:

1). The Brazilian Amazon covers around 4 million square kilometers (almost half the national territory), representing 69% of the Amazon basin. Annual rainfall is generally above 2000 mm and decreases from the equatorial regions towards the tropics and the Northeast of Brazil (under 1500 mm).

2). Caatinga is described as the most biodiverse and the most populated semi-arid region in the world (MMA, 2011). It mostly receives less than 750 mm rainfall year-1 (Leal et al., 2005), with peaks in Mach-April over the north and the center part of the region and in November-March over the southern part. The year to year rainfall variability can be greater than 40 % (Moura and Sukla, 1981).

3). Cerrado is a tropical savanna covering 22% of Brazil's territory. The climate is punctuated by a severe dry season that ranges between three and five months from May to September. The overall amount of rain is higher with 800-2000 mm year-1 (Ratter et al., 1997).

4). Atlantic Forest is the second-largest rainforest of the American continent and one of the world's 25 biodiversity hotspots. This region receives from 1000 to 3000 mm annual rainfall.

5). The Brazilian Pampa represents 2.07% of the national territory and lies within the South Temperate Zone (Roesch et al., 2009). The annual precipitation in the Pampean region is around 1,200–1,600 mm.

6). The Pantanal wetland is a vast complex of seasonally inundated floodplains along the upper Paraguay River, located mostly in Brazil (Hamilton et al., 2002). Annual rainfall is 1000-1500 mm across the basin, with most rainfall occurring between November and March.

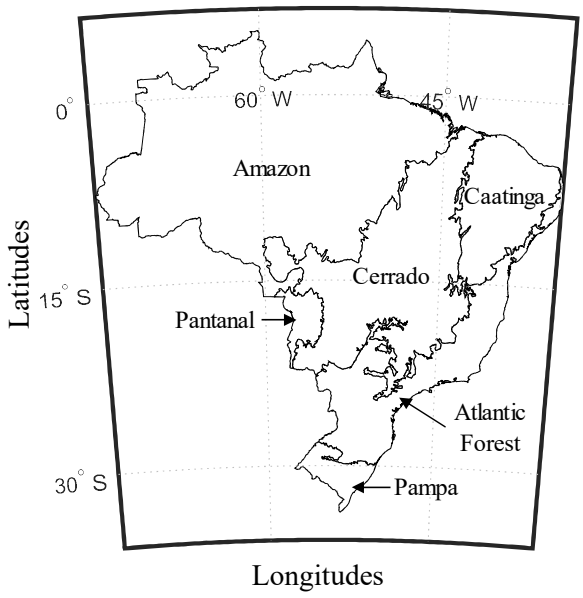

Figure IV.1. Regions of Brazil involved in this study corresponding to the six major natural biomes as defined in IBGE (2016).

2.2 Verification dataset

The choice of the verification dataset is important in the context of medium-range forecasting, especially in data-sparse regions affected by complex patterns of variability. Using gridded data based on rain gauge observations has the advantage of being independent of all models (Hagedorn et al., 2012). Carvalho et al. (2012) found that the Physical Science Division South America Daily Gridded Precipitation dataset (Liebmann, and Allured, 2005; Liebmann, and Allured, 2006) consistently represents the variability of the South American monsoon system, which is the most important climatic feature in South America and provides a similar spatial pattern of mean precipitation compared with other gridded precipitation products such as the Global Precipitation Climatology Project (Huffman et al. 2001) and Climate Prediction Center unified gauge (Silva et al. 2007). This dataset has been constructed using historical records from rain gauge stations. In this study, we use this dataset for evaluating rainfall forecasts over each biome in Brazil. The verification dataset consists of $1^{o}\times1^{o}$ grid values of daily precipitation over Brazil over 1985-2010, which was interpolated using the average precipitation within a geographic ellipse. Measurements have been taken at 1200 UTC, while precipitations are recorded as having occurred on the day on which the rain gauge reading is taken. This dataset is available at http://www.esrl.noaa.gov/psd/data/gridded/data.south_america_precip.html. Figure IV.2 shows the cumulative probabilistic distribution of the daily precipitations over each biome generated from the verification dataset.

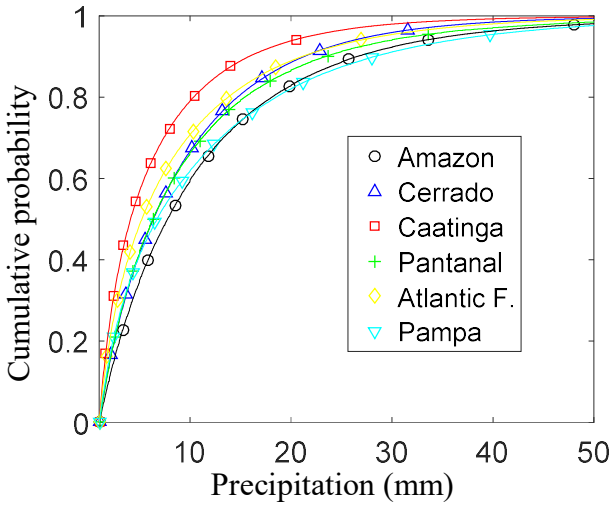

Figure IV.2. Cumulative distribution of the precipitations higher than 1 mm over 1985-2010.

It is worth noting that at least two other datasets based on rain gauge observations are available for Brazil. Silva et al. (2007) produced the Climate Prediction Center unified gauge. This is a 1° × 1° degree dataset using a Cressman (1959) scheme of interpolation (Glahn et al. 1985) that corrects the background gridpoint value by a linear combination of residuals between calculated and observed values. However, this dataset has fewer rain gauges over the Brazilian Amazonian domain compared to the adopted dataset (see Fu et al., 2013). Recently, Xavier et al. (2016) produced a high-resolution dataset over a 0.25∘ × 0.25∘ grid based upon the inverse distance weighting interpolation method; this method had been identified as the most skillful when compared against several other methods. However, the grid coordinates in this dataset do not coincide with the grid coordinates of the forecast datasets in our study, meaning that a further interpolation would be needed to use it as our verification dataset.

2.3 Forecast datasets

2.3.1 Global Ensemble Forecast System (NCEP) reforecast data

We used 1º×1º gridded reforecasts produced from the 2nd-Generation Global Medium-Range Ensemble Reforecast Dataset (Hamill et al., 2013). This a retrospective weather forecast dataset generated with the currently operational NCEP Global Ensemble Forecast System (NCEP), available at http://esrl.noaa.gov/psd/forecasts/reforecast2/download.html.

Our daily precipitation ensemble reforecasts considered both, the control forecast and the 10 perturbed forecasts, of cumulative six-hours total precipitations issued at 00 UTC hours over 1985-2010 (26 years) at 1.5-, 3.5- and 5.5-day leads. A lead time of 1.5 days matches up the observation of day *n* with the sum of the six-hours total precipitation at 18, 24, 30 and 36 hours of the forecast issued at day *n*-1. Probabilities were calculated directly from the ensemble relative frequency, referred to as ‘‘raw’’ probabilities henceforth.

2.3.2 ECMWF forecasts data

ECMWF reforecasts archived in the TIGGE database at ECMWF (see http://apps.ecmwf.int/datasets/data/tigge) were also considered. We used the 50 member ensembles of perturbed ECMWF forecasts issued at 00 UTC hours over October, 2006 - 2010 at lead times of 1.5, 3.5 and 5.5 days. Forecasts were bilinearly interpolated into a 1º latitude-

longitude grid using ECMWF's TIGGE portal software. About 2.0 percent of the records accounted for negative, mostly negligible values, that were set to zero. Probabilities were also calculated directly from the ensemble relative frequency

2.4 Post-processing methods

2.4.1 The analog forecast method

In the analog forecast method, the real-time forecast is adjusted using a long time series of past forecasts and associated observations (Hamill et al., 2015). Suppose that we want to produce an ensemble of *n* analog forecasts for today's forecast at a specific point and a given lead. The first step is to compare the today's forecasts within a region surrounding that point with the forecasts from the historical reforecast archive in that same region and at the same forecast lead, and then find the *n* dates with the best matching. In a second step, the analog ensemble is formed from the verification dataset on those dates. This process is repeated for each lead day and location across the study region, and the forecast over the entire region is produced by tiling together the local analog forecasts (Hamill et al., 2006; Tian and Martínez, 2014). Leave-one-out cross-validation is carried out by excluding the current year from the list of potential analogs. For a detailed description and theoretical basis of the analog method, the readers can refer to Hamill and Whitaker (2006).

2.4.2 Logistic regression method

In the logistic regression method a nonlinear function is fitted to past pairs of the predictor(s), and the predictand, which as an observed value takes on a probability of either 1.0 (the event occurred) or 0.0 (the event did not occur), according to the adopted threshold *T* (Wilks 2006). The fitted function is then used to estimate the probability *P* that the current unknown observed amount *O* be higher than the threshold *T* given the current predictor values, associated with the forecast. In this study we adopted the same nonlinear function as Hamill et al. (2008):

$$P(O > T) = 1 - 1/\left(1 + \exp\left(a + b\bar{F}_{pr}^{0.25} + c\sigma_{F_{pr}}^{0.25}\right)\right) \qquad \text{(IV.1)}$$

where $\bar{F}_{pr}$ and $\sigma_{F_{pr}}$ represent the mean and the standard deviation of the ensemble of precipitation forecast, respectively, while $a, b,$ and c are the fitting parameters. Following Hamill and Whitaker

(2006), we also pondered using a one-half power transformation of the predictors, instead of the one-quarter adopted here, but our results were practically the same.

As for the analog method, the logistic regression technique is performed separately for each location and each forecast lead time, within the verification period using all historical data available.

2.5 Experimental design

The first experiment is to compare the performance of NCEP and ECMWF raw forecasts, as well as NCEP analog forecasts over January, April, July and October, from October 2006-December 2010. These four months are representative of the summer, fall, winter and spring season, respectively. In this and the subsequent experiments, training of the NCEP forecasts considered the 26-years dataset of retrospective forecasts. The analog forecasts for current date and time were formed by finding the lowest sum of the square differences (ssd) between the current forecasts and the similar historical forecasts in the other years (25 in total) from the reforecast archive, considering a limited region of 9 grid points. The forecasts were selected within a ±45-day window around the date of the forecast and the best 50 analogs were chosen to construct the forecast ensemble. This analog procedure is adopted as the control variant of the method and referred henceforth as the "Control" forecast.

The second experiment is to conduct an inter-comparison among six NCEP-based analog approaches and one logistic regression method. In this case, the forecasts were verified over January, April, July and October from 1985-2010. The six analog-based methods included the Control method plus five modified versions of this procedure (see Table IV.1), with each version considering only one modification with respect to the Control procedure. Each method is described as follows:

- Short_reg considered a search region with five grid points, i.e., the current grid point and the four adjacent grid points at a distance of 1°.
- 100_Ens was produced with 100 analog members, instead of only 50.
- LogF considered the differences between the logarithm of the current and past precipitation forecasts plus one, as the measure of the closeness among forecasts.

- A_09pr_01pw included the mean ensemble of the column precipitable water as a predictor variable. The analogs were produced by pondering the 90% of the ssd of total precipitation plus 10% of the ssd of precipitable water.
- A_05pr_05pw is similar to A_09pr_01pw, but considering the 50% of ssd of both, total precipitation and precipitable water.

Besides these five methods, we had pondered the rank analog technique (Hamill and Whitaker, 2006), which used the differences between the rank of the precipitation forecasts within the search region as the similarity measure. However, this method was found unsuitable for the conditions of Brazil and therefore excluded.

Table IV.1. Configurations of the six analog approaches

ID_Method	Ens. size	Grid points	Closeness metric
Control	50	9	$\sum_{i=1}^{N_s+1}\left(F_{pr}^{i,t}-F_{pr}^{i,tc}\right)^2$
Short_reg	50	5	$\sum_{i=1}^{N_s+1}\left(F_{pr}^{i,t}-F_{pr}^{i,tc}\right)^2$
100_Ens	100	9	$\sum_{i=1}^{N_s+1}\left(F_{pr}^{i,t}-F_{pr}^{i,tc}\right)^2$
LogF	50	9	$\sum_{i=1}^{N_s+1}\left(\log\left(F_{pr}^{i,t}+1\right)-\log\left(F_{pr}^{i,tc}+1\right)\right)^2$
09pr_01pw	50	9	$0.9\times\sqrt{\sum_{i=1}^{N_s+1}\left(F_{pr}^{i,t}-F_{pr}^{i,tc}\right)^2}+0.1\times\sqrt{\sum_{i=1}^{N_s+1}\left(F_{pw}^{i,t}-F_{pw}^{i,tc}\right)^2}$
05pr_05 pw	50	9	$0.5\times\sqrt{\sum_{i=1}^{N_s+1}\left(F_{pr}^{i,t}-F_{pr}^{i,tc}\right)^2}+0.5\times\sqrt{\sum_{i=1}^{N_s+1}\left(F_{pw}^{i,t}-F_{pw}^{i,tc}\right)^2}$

$F_{pr}^{i,t}$ and $F_{pw}^{i,t}$ are the 24 hours cumulative precipitation pr and the total-column precipitable water pw forecasts, respectively, at time t and over grid point i, while $F_{pr}^{i,tc}$ and $F_{pw}^{i,tc}$ are the corresponding forecasts at current time tc. (involving the current grid point and the set of N_s supplemental points surrounding the current grid point) and time t.

By matching ranks instead of the actual values, many members of the analog ensemble corresponded to dates whose precipitations over the search region follows the same order (rank) compared to the current day, but whose total amounts are dramatically different. For example, the method often matched a heavy rainfall at the current day with a drizzle in the past.

2.6 Verification analysis

In this study we compare point and regionally aggregated values (see Medina et al., 2018) of several deterministic and probabilistic metrics. For the deterministic metrics, we used the mean

error (ME) the root mean square error (RMSE) and correlation coefficient (ρ), which are among the most commonly reported measures of agreement between forecasts and observations. For the probabilistic metrics we used the Brier Skill Score (BSS) and the reliability diagram (Wilks, 2011) associated with the precipitation events above 2.5 mm. In the study the forecast probability is calculated from the ensemble forecast, while the climatological probability is computed as an average probability taken over ±30 days of the forecast date. A bootstrapping procedure involving 1000 samples was used to quantify the uncertainty of the probabilistic statistics (see Medina et al., 2018).

3 RESULTS AND DISCUSSION

3.1 Inter-comparisons between NCEP, ECMWF, and Control analog post-processed forecasts

Figure IV.3 shows the average correlation and RMSE of the raw NCEP forecasts in each region, and their differences with the ECMWF forecasts and the Control analog forecasts.

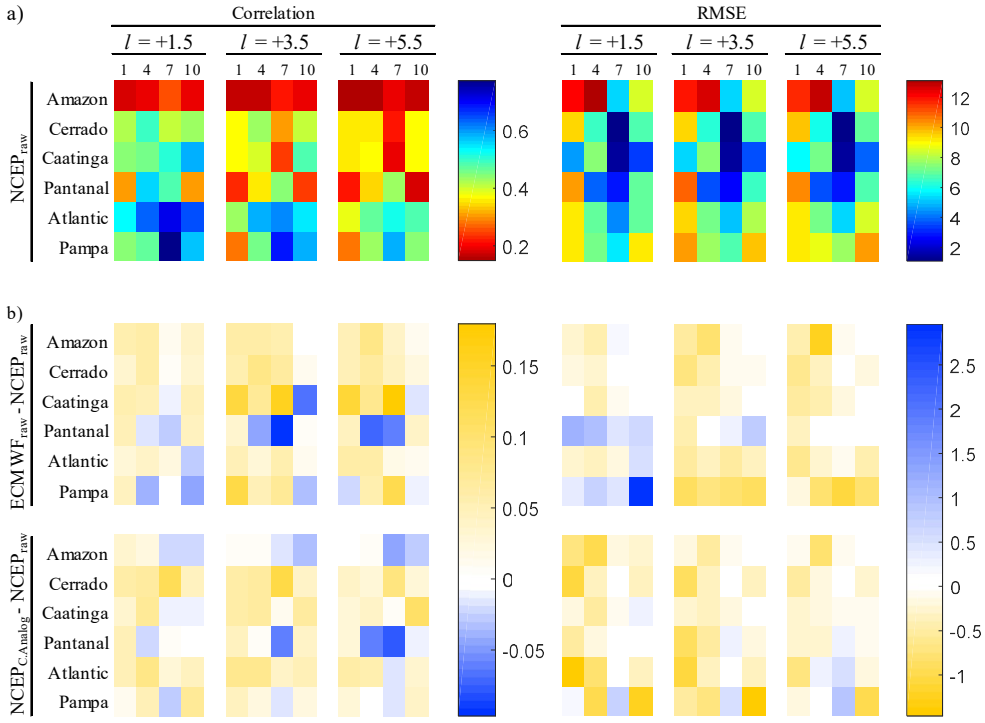

Figure IV.3. a) correlation and RMSE (mm) of the NCEP precipitation raw forecasts at each biome in January (1), April (4), July (7) and October (10), for lead times 1.5, 3.5 and 5.5; and b) differences between correlation and RMSE of the ECMWF raw forecasts as well as the Control analog forecasts and the NCEP raw forecasts.

The average correlation varied especially among regions: from high values over Atlantic Forest and Pampa, to very weak values over Amazon. The RMSE was proportional to the total rainfall and therefore more seasonally driven, with maximums during warm seasons, and minimums during cold seasons.

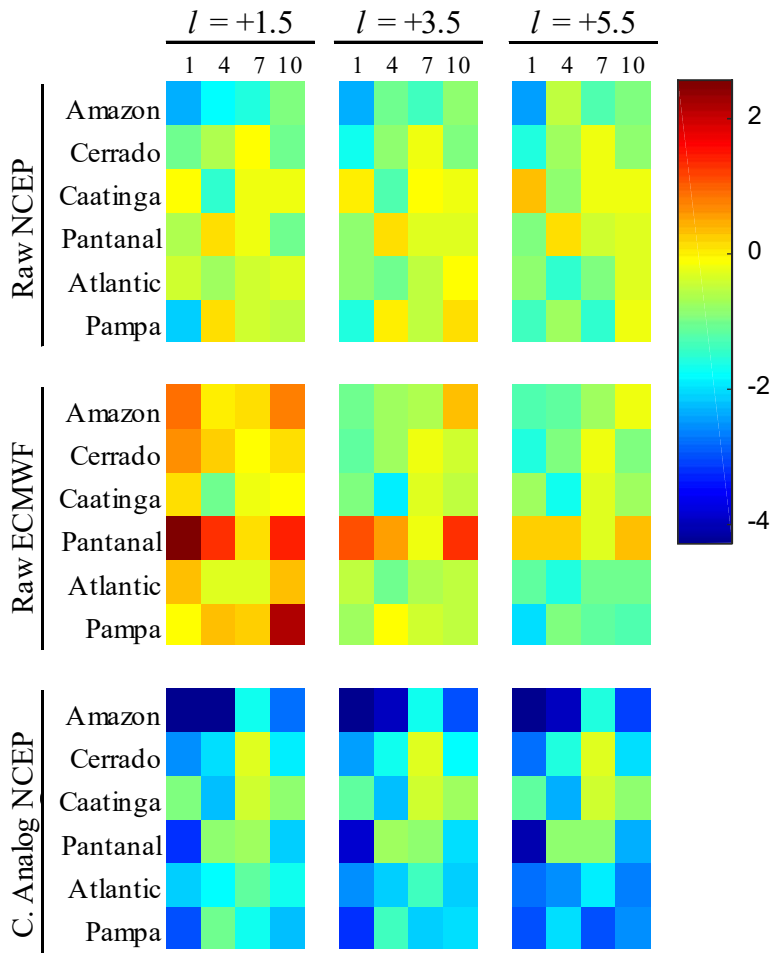

Figure IV.4. ME (mm) of the raw NCEP and ECMWF and the Control analog precipitation forecasts at each biome in January (1), April (4), July (7) and October (10), for lead times 1.5, 3.5 and 5.5.

Janowiak et al. (2010) noted a very weak correlation between the ECMWF and NCEP forecasts and the Global Precipitation Climatology analyses over the Northwest of South America in warm seasons. The ECMWF and NCEP raw forecasts performed comparably at 1.5 days, but the former performed better, even compared with the Control analog forecasts, at 3.5 and 5.5 days. The performance of the NCEP forecasts mostly improved through post-processing; the analog control forecast provided the best correlation and RMSE at 1.5 days. Moreover, as indicated in Figure IV.4, the Control analog NCEP forecasts showed greater ME than both NCEP and ECMWF raw forecasts. They tended to underestimate the precipitations in most regions, seasons, and lead times.

Figure IV.5 presents the distribution of the bootstrapped BSS values over each region and month at lead times 1.5 and 5.5-day. The Control analog forecasts in most cases improved the BSS compared to the raw ensemble forecasts.

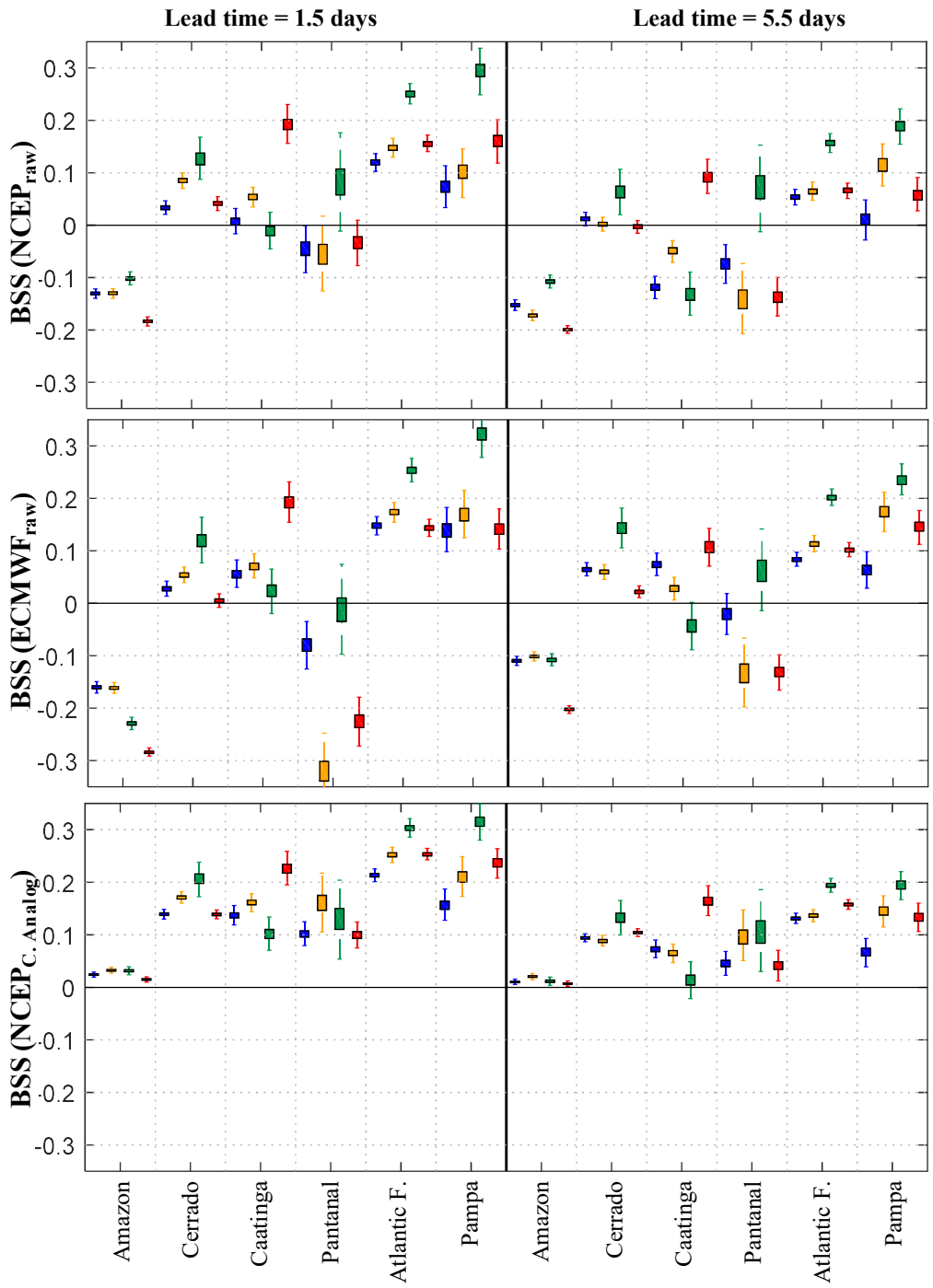

Figure IV.5. BSS of the raw NCEP and ECMWF and the Control analog precipitation in (from left to right) January (blue), April (yellow), July (green), and October (red) for lead times 1.5 and 5.5-day.

The improvements tend to be higher over regions and seasons, such as the spring month in Amazon and the fall month in Pantanal, where raw forecasts are less skillful. Practically all the Control analog forecasts provided a positive BSS, although it was still close to zero in Amazon.

The raw ECMWF forecasts showed higher probabilistic forecast skill than the raw NCEP forecasts at 3.5 and 5.5 lead days, but lower at 1.5 days. Both the ECMWF and NCEP raw forecasts showed no skill over Amazon and Pantanal at any lead time, indicating that the climatological predictions are better here compared to the raw forecasts.

This result is consistent with the study based on regional ensemble forecasts over South America (Ruiz et al., 2009). The reason for that is due to the Convection in the Amazon exhibiting more pronounced diurnal and seasonal variability than in the East region (Jones and Schemm, 2000). To provide a better insight in space, Figure IV.6 shows the differences between the Brier Scores of the raw ECMWF and NCEP ensemble forecasts at each grid point. Positive differences indicate the NCEP forecasts are better, since the lower the Brier Score the better the forecasts.

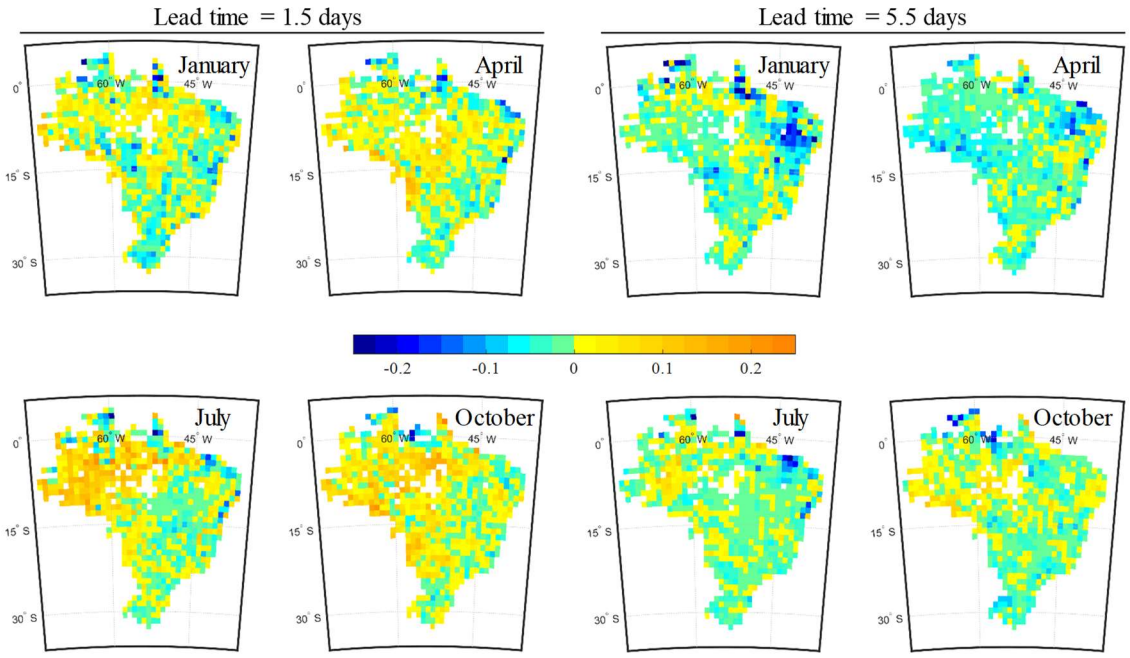

Figure IV.6. Differences between the ECMWF and the NCEP Brier_score for 1.5 and 5.5 lead days.

The ECMWF forecast seems relatively weak over the Northwest, mainly at 1.5 days, probably due to issues with the model representation of the daily precipitation cycle over Amazon (Betts and Jakob, 2002a,b). Similarly, the NCEP forecasts are unskillful over this region as well. Seasonally, the NCEP forecasts tend to be relatively better in October (the spring month), a period associated with the onset of the precipitations in most Brazil (Marengo et al., 2001, Grimm and Zilly, 2009), and the ECMWF forecasts in January (a period of higher convection), at both lead times. In the contrasting ECMWF and NCEP forecast performance among lead times mediated the

fact that the ECMWF forecasts at 3.5 and 5.5 days in several cases provided better RMSE and BSS than the forecasts at 1.5 days; the bias of the ECMWF at 5.5 days tend to be negative while the bias at 1.5 days is positive (Fig. IV. 4). These trends were not observed for the raw and post-processed NCEP forecasts. This finding is consistent with Janowiak et al. (2010), who found that the 9-day ECMWF raw forecasts had lower bias than the day-2 forecasts over much of central South America. To investigate what caused the better performance in longer lead time, we analyzed the spread-skill relationships of different forecasts at 1.5 and 5.5 leads, by comparing the average standard deviation of the ensembles to the RMSE of the ensemble means for different intervals of the deviations (Figure IV.7). The result showed that, while the spread of the NCEP ensembles (both, raw and post-processed) was similar at different lead times, the ECMWF ensembles at 1.5 lead days were more underdispersed than 5.5 lead days. The wider spread of the ECMWF ensemble forecasts at longer lead times compared to shorter lead times may cause the better performance for longer lead times. The results also suggest that the forecast post-processing with the analog technique considerably improves the spread-skill relationship of the ensembles.

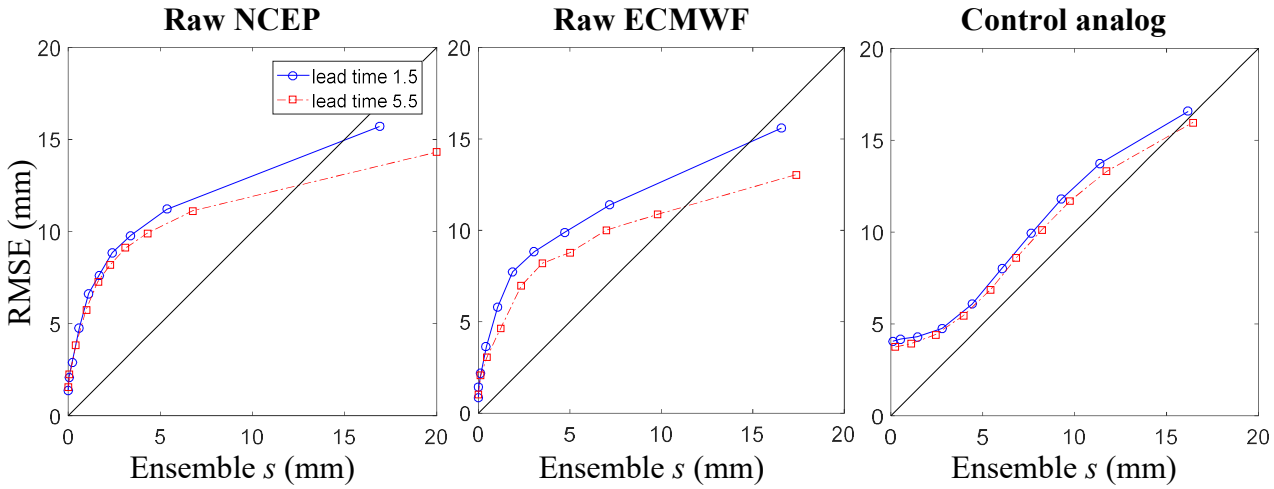

Figure IV.7. RMSE of the ensemble forecasts versus the mean standard deviation *s* of the ensemble members over all grid points and at +1.5 and 5.5 lead days, from left to right, raw NCEP, raw ECMWF, and Control analog forecasts.

Figure IV.8 shows the reliability diagrams over January at lead days 1.5 and 5.5. In general, the forecasts were slightly less reliable in drier months when high probability precipitation forecasts are issued less frequently. The post-processed forecasts were considerably more reliable but less sharp than raw forecasts. The frequency of medium-probability forecasts grows after post-processing mainly at the expense of the high-probability forecasts, as found by Hamill et al. (2008).

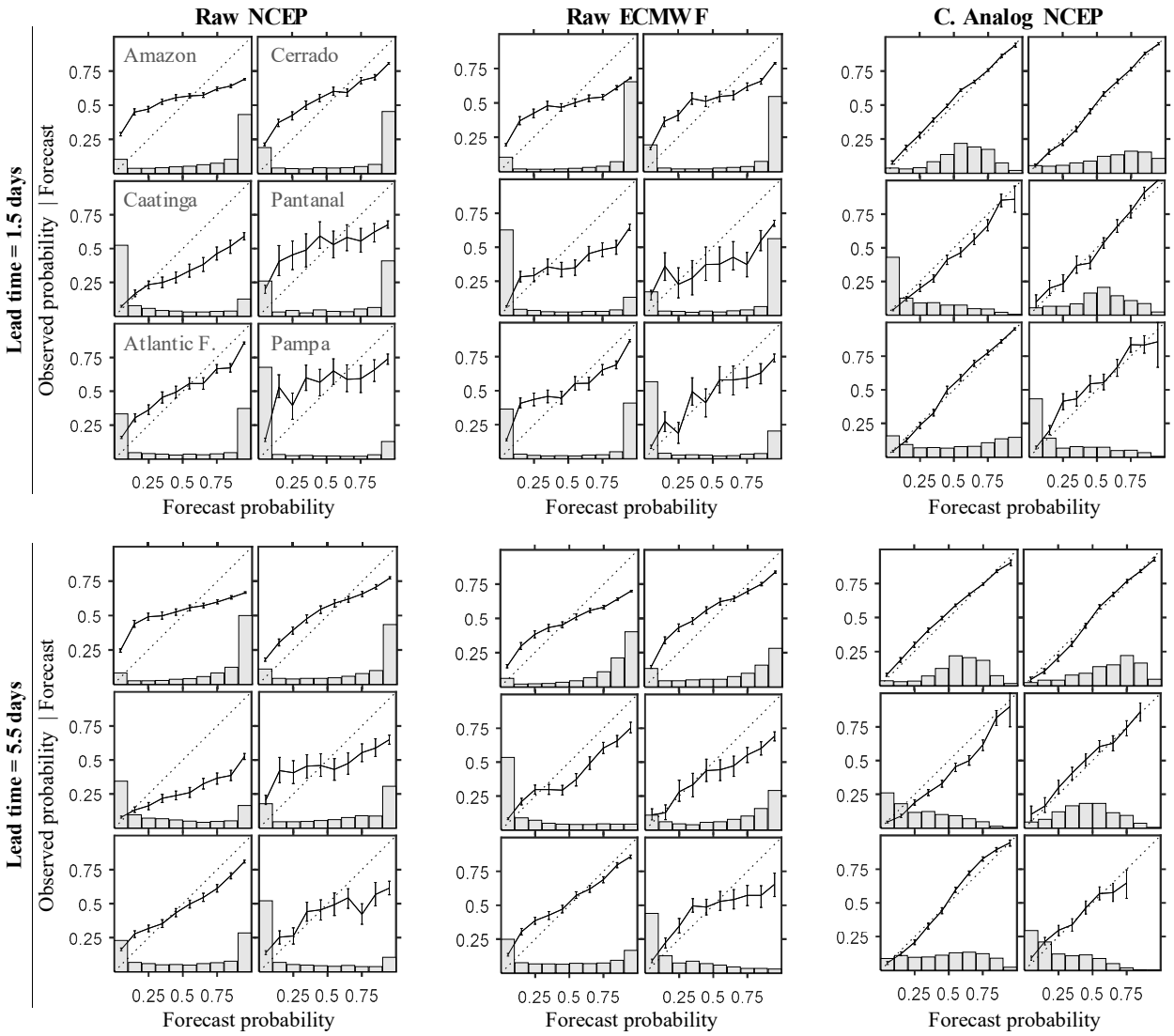

Figure IV.8. Reliability diagrams of the NCEP and ECMWF raw forecasts, and the Control analog forecasts for January for days +1.5 and +5.5.

The NCEP and ECMWF raw forecasts provided similar reliability at 1.5 lead days, while the former one seemed slightly less reliable at 5.5 days. While the reliabilities are not considerably changed with lead times, the sharpness at shorter lead times is slightly higher than longer lead times, especially for the ECMWF forecasts. This may be caused by the narrower ensemble spread at shorter lead times. It is also worth noting that through post-processing, the reliability of the precipitation forecasts improved more than the skill score, which is in agreement with previous studies based on, either analog post-processing techniques (e.g., Voisin et al., 2010) or other methods (Hamill et al., 2012).

In summary, the Control analog forecasts considerably improved the probabilistic forecasting performance but more systematically biased compared to the NCEP and ECMWF raw forecasts. They were also slightly less correlated with observations and less accurate than the ECMWF forecasts at 3.5 and 5.5 days. The performance of the raw ECMWF and NCEP forecasts was comparable at the 1.5-day lead, but the ECMWF forecasts performed better at the longer lead times.

3.2 Comparing multiple analog approaches and the logistic regression approach

Figure IV.9 shows cumulative probability distributions of the correlations, the ME and the RMSE using raw NCEP forecasts and the six analog approaches over January and at 1.5 days lead. The trends were similar across different regions and lead times. While all the analog forecasts considerably improve the correlation and the RMSE compared to the raw forecasts, they are more systematically biased than the raw forecasts. This may reflect issues of the analog procedures to find an appropriate number of analogs when the current precipitation forecast is especially large (Hamill, 2015).

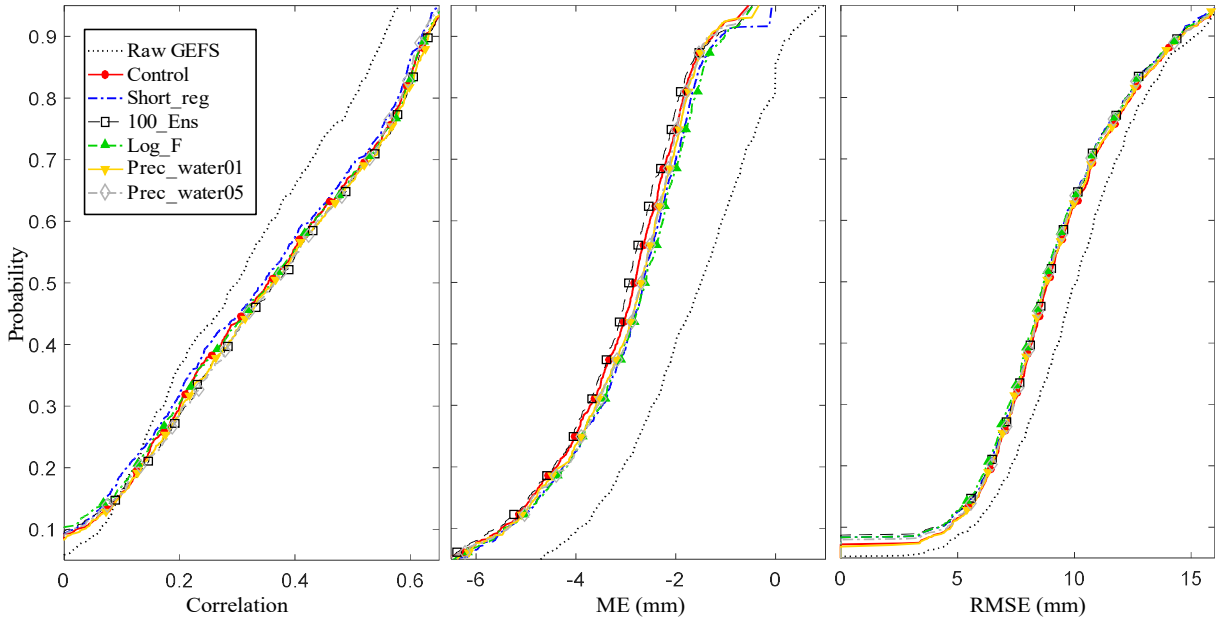

Figure IV.9. Cumulative distribution of the correlations, ME and RMSE for the raw and analog NCEP forecasts at 1.5 lead days in January.

A paired sample t-test was conducted to compare the performance of the six analog forecasts. The result shows that the differences of the six analog methods are small in correlation

and ME but mostly significant at the 1% significance level; the differences in RMSE values are less significant, especially for comparisons among the Control, Sort_reg, and 09pr_01pw approaches. The changes in ME and RMSE after post-processing seemed roughly constant among grid points, while the correlation improved more over grid points with higher correlations, i.e., regions with better correlation were more benefited through post-processing. Among all the six methods, the 100_Ens and LogF forecasts commonly provided the best correlations and ME and RMSE, respectively. The 05pr_05pw forecasts in most cases perform the worst among all the analog forecasts.

For probabilistic forecasts, all the analog forecasts, as well as the LR forecasts considerably improved the BSS compared to the raw forecasts (Figure IV.10).

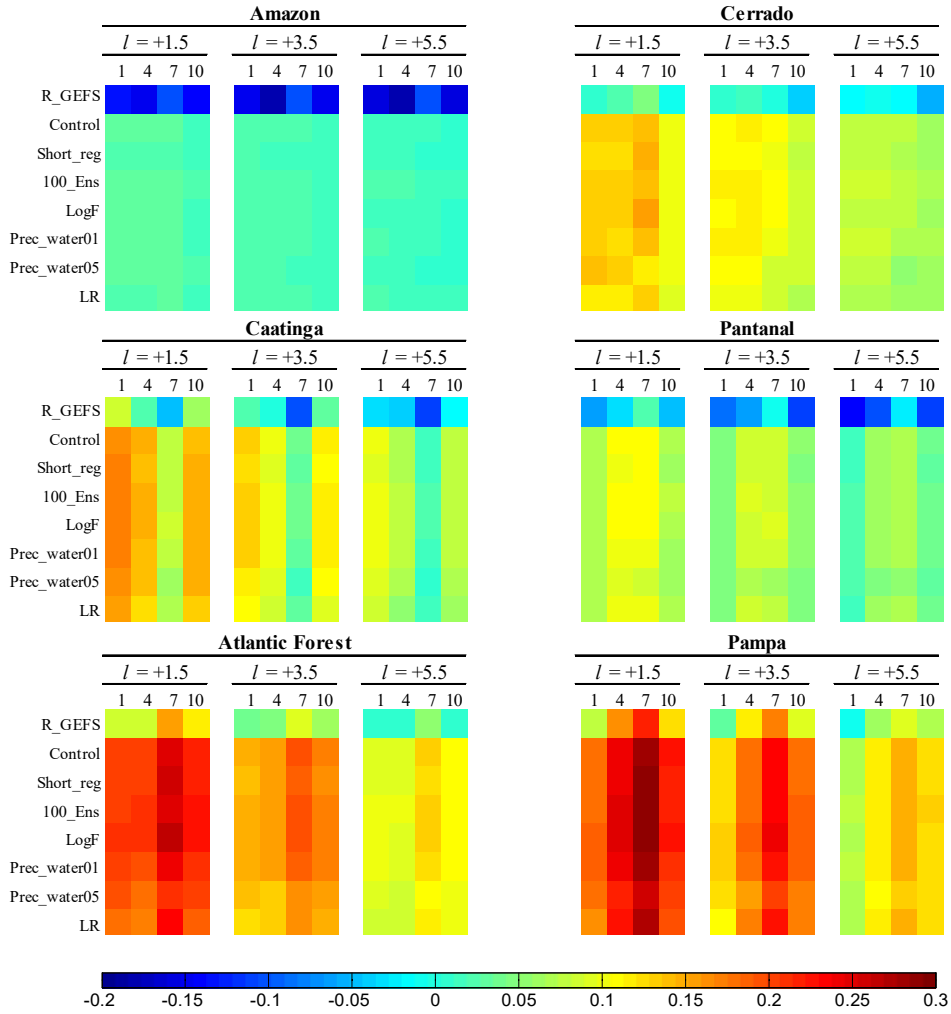

Figure IV.10. Mean BSS of the raw NCEP forecasts, the analog calibration methods and the logistic regression method in the six regions.

The improvements were similar but mostly significantly different at the 1% significance level. In agreement with previous studies (e.g., Hamill and Whitaker, 2006, Delle Monache et al., 2013), the forecasts produced with analog methods provided better skill compared to those produced with the logistic regression. Only the 05pr_05pw forecasts performed similarly or slightly worse than the LR forecasts. While the 100_Ens forecasts commonly provided better BSS over West regions, where the skill is consistently low, the Log_F forecasts provided better skill over the East regions. The averages BSS were mostly below 0.3 and affected by the considerable spatial and temporal variability of the BSS (Figure IV.11). As suggested by the maps of Brier Score of the climatology in Figure IV.12, this variability seemed associated with the climate predictability.

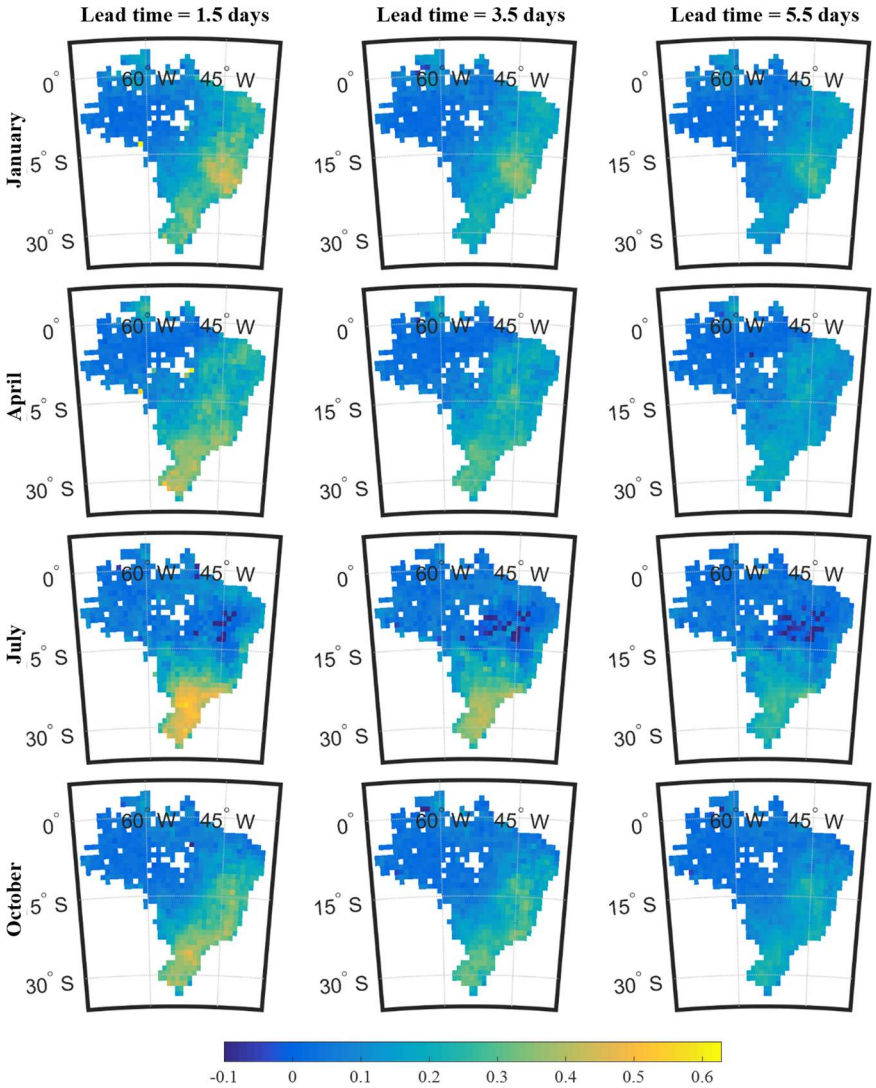

Figure IV.11. BSS values of the basic analog technique in space from 1985-2010.

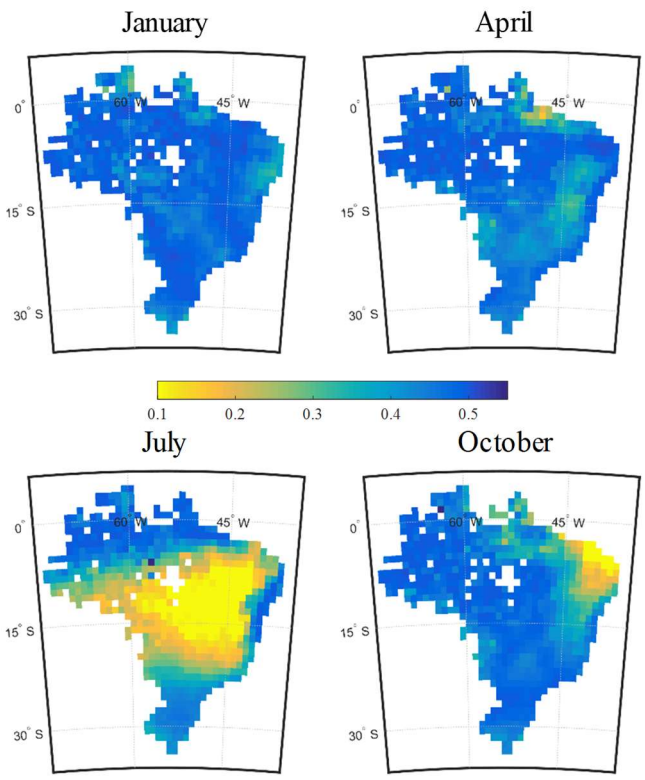

Figure IV.12. Brier score of the climatology in space. 2.5mm is used as a threshold

Climate predictability changes from high in the winter all over the center of Brazil, to very low in summer practically all over the country. It is influenced by the interannual migration of deep tropical convection from the central and southern portion of the Amazon basin in summer to the northwestern sector of South America in winter (Rao and Hada, 1990).

Figure IV.13 shows the reliability diagrams of the raw and post-processed forecasts over January at 1.5 days lead. The post-processed forecasts performed very similarly: they were considerably more reliable but less sharp than the raw forecasts. Only the 100_Ens forecasts showed a slightly dry bias over a few regions. Our post-processed forecasts are slightly more reliable than the analog forecasts from Hamill and Whitaker (2006) and the logistic regression forecasts from Hamill et al. (2008), probably due to the new improvements of the NCEP model compared to its first version (Hamill et al., 2015).

3.3 Discussing analog post-processing methods for precipitation forecasts

While the comparison between raw and post-processed forecasts in section 3.1 only involved one analog approach, the results from this comparison seemed also valid to the other post-processing methods. This is because the differences among post-processing methods were lower than the differences between raw and post-processed forecasts. Nonetheless, different methods in most cases provided significantly different statistics, allowing us to identify the best and the worst strategy. In Table IV.2 we show the number of experiments (of a total of 72, i.e., 6 regions × 4 months × 3 lead times) where each analog approach performed best and worst in terms of correlation, ME, RMSE and BSS. We also provide a ranking of the methods by sorting the

differences between totals, from better to worse. The LR based performance was considered when comparing the BSS, but not ranked.

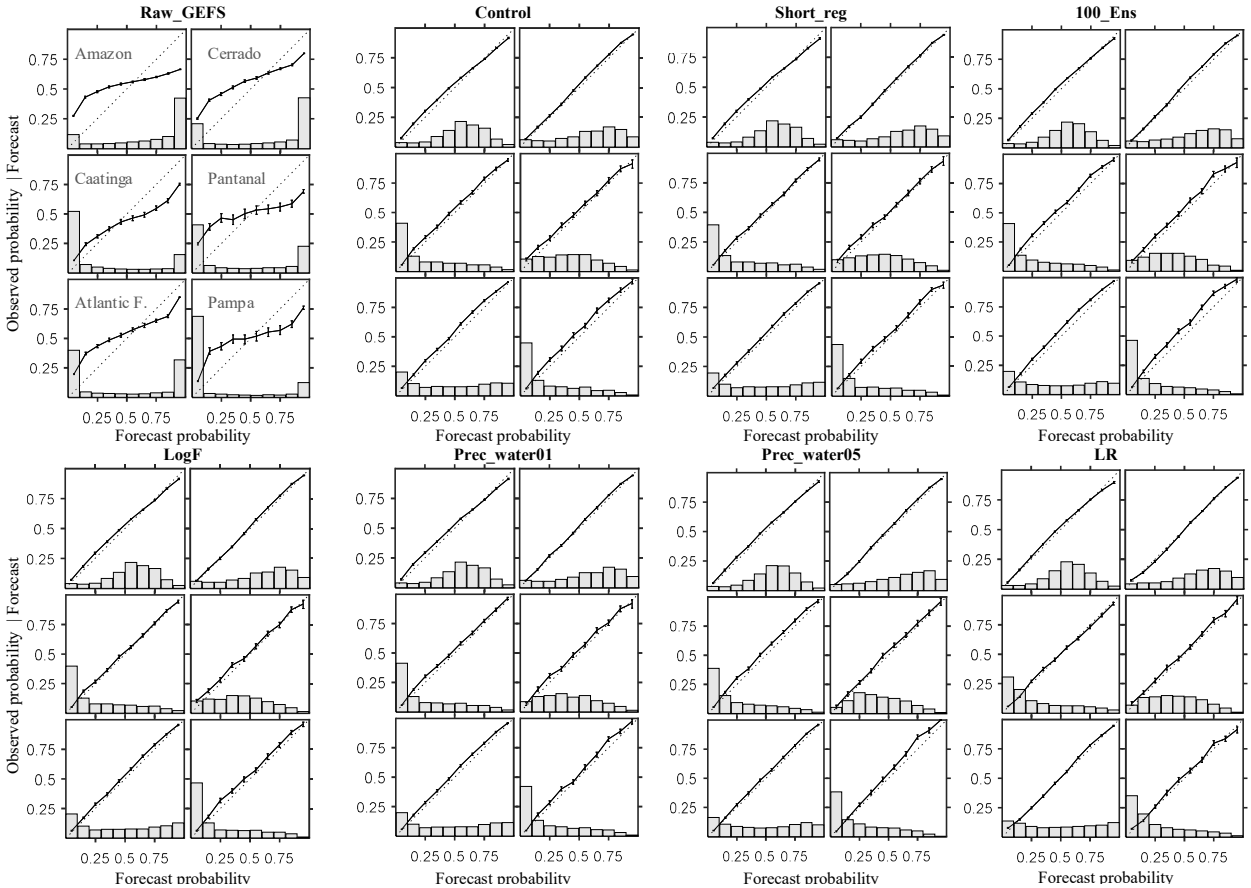

Figure IV.13. Reliability diagrams for the NCEP-based precipitation forecasts in January. Each panel indicates each reliability diagram of the methods including raw NCEP, six analog methods, and logistic regression.

LogF was identified as the best approach, followed by the 100_Ens approach. Finally, the 05pr_05 pw forecasts performed worst for most statistics, followed by Short_reg. The LogF forecasts ranked as the best or the second-best in terms of all the performance statistics. The 100_Ens forecasts provided better BSS and correlations but worse ME and RMSE. The larger the size of the ensemble, the more the difficulty for identifying less unbiased members on days with large precipitations, especially for regions with lower pluviometry. These results led to the question of whether an analog method combining the modifications adopted in Log_F and 100_Ens (100Ens-LogF) improves the performance compared to each of these two approaches. We found that it slightly improved the performance of the original approaches by taking advantage of their best features. In particular, it provided consistently better BSS than the Log_F and 100_Ens

strategies, although the improvements only affected the third significant figure. It also improved the ME compared with the 100_Ens strategy, but not compared with the Log_F strategy.

Table IV.2. Number of experiments (considering 6 regions, 4 months and 3 lead times) where the alternative analog approaches performed the best and worse in terms of different metrics.

Method	Correlation		ME		RMSE		BSS		Ranking
	Best	Worst	Best	Worst	Best	Worst	Best	Worst	
Control	2	3	0	3	2	1	1	1	4
Short_reg	1	22	3	2	2	4	0	9	5
100_Ens	57	1	0	33	8	16	37	1	2
LogF	7	1	60	0	48	2	24	0	1
09pr_01pw	3	1	4	1	11	4	5	0	3
05pr_05 pw	2	44	5	33	1	45	3	28	6
LR							2	33	

4 CONCLUSIONS

This study conducted inter-comparisons between raw NCEP forecasts, raw ECMWF forecasts, and post-processed NCEP forecasts with six analog methods and the logistic regression method over six biome regions in Brazil. To the authors' knowledge, this study is the first to comprehensively examine the performance of these global-scale NWP models and statistical post-processing methods over South America, specifically over a region severely affected by large mesoscale convective systems.

The article showed that the global scale NWP's raw forecasts are helpful for precipitation forecasting over the East, and particularly the Southeast, of Brazil, but unskillful over the Northwest. The ECMWF raw forecasts are better than the NCEP raw forecasts since they perform similarly or better over the East. However, the post-processed NCEP forecasts, particularly the analog forecasts, are strongly recommended over the raw ECMWF forecasts as they performed probabilistically much better; unlike the raw forecasts they improved the skill of climatological forecasts in all the evaluated regions, seasons and lead times. Our results also confirmed previous findings showing that the analog forecasts tend to be negatively biased: this study suggests that the larger the size of the analog ensemble the higher the bias.

The forecast performance showed less sensitivity to the post-processing strategy than to the post-processing itself. Nevertheless, different post-processing strategies are significantly

different statistically, with the analog forecasts being as reliable as the logistic regression forecast but slightly more skillful. The strategy considering the log of current and past reforecasts as the measure of closeness performed slightly better among all the analog forecasts, followed by that considering 100 analog members (instead of the regular 50). The analog method combining modifications adopted in these two approaches performed slightly better than the individual approaches. Whereas, the strategies that included precipitable water as a predictor variable were among the worse.

This study provides useful information for precipitation forecasting over tropical and subtropical regions affected by large mesoscale convective systems. While we have addressed the impact of the forecast uncertainty on the performance by using bootstrap analysis, we have not addressed the impact of the verification dataset uncertainty. The quality of interpolated datasets in data-sparse regions is always a source of concern. Neither do we analyzed whether the combination of the ECMWF and NCEP raw forecasts provide any improvement compared to the single model forecasts. More research is still needed for further decreasing the forecasting uncertainty, especially over the Amazon. We foresee future studies will evaluate the efficacy of multimodel forecasts and other post-processing methods with the consideration of the uncertainty from the verification dataset, in particular focusing on the methods that can perform well with much shorter training data sets.

REFERENCES

1. Bauer, P., Thorpe, A., Brunet, G. 2015. The quiet revolution of numerical weather prediction. Nature, 525(7567), 47-55.
2. Bechtold, P., Bauer, P., Berrisford, P., Bidlot, J., Cardinali, C., Haiden, T., Janousek, M., Klocke, D., Magnusson, L., McNally, A., Prates, F., 2012. Progress in predicting tropical systems: The role of convection. European Centre for Medium-Range Weather Forecasts.
3. Bechtold, P., Semane, N., Lopez, P., Chaboureau, J.P., Beljaars, A., Bormann, N., 2014. Representing equilibrium and nonequilibrium convection in large-scale models. Journal of the Atmospheric Sciences, 71(2), 734-753.
4. Betts, A.K., Jakob, C., 2002a. Evaluation of the diurnal cycle of precipitation, surface thermodynamics and surface fluxes in the ECMWF model using LB A data. Journal of Geophysical Research: Atmospheres, 107(D20), doi:10.1029/2001JD000427.

5. Betts, A.K., Jakob, C., 2002b. Study of diurnal cycle of convective precipitation over Amazonia using a single column model. Journal of Geophysical Research: Atmospheres, 107(D23), doi:/10.1029/2002JD002264
6. Bony, S., Stevens, B., Frierson, D.M., Jakob, C., Kageyama, M., Pincus, R., Shepherd, T.G., Sherwood, S.C., Siebesma, A.P., Sobel, A.H., Watanabe, M., 2015. Clouds, circulation and climate sensitivity. Nature Geoscience, 8(4), 261.
7. Buizza, R., Houtekamer, P.L., Pellerin, G., Toth, Z., Zhu, Y., Wei, M., 2005. A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. Monthly Weather Review, 133(5), 1076-1097.
8. Carvalho, L.M., Jones, C., Posadas, A.N., Quiroz, R., Bookhagen, B., Liebmann, B., 2012. Precipitation characteristics of the South American monsoon system derived from multiple datasets. J. Climate, 25(13), 4600-4620.
9. Charba, J.P., Harrell, A.W., Lackner, A.C., 1992. A monthly precipitation amount climatology derived from published atlas maps development of a digital database.
10. Cloke, H.L., Pappenberger, F., 2009. Ensemble flood forecasting: a review. J. Hydrol., 375(3-4), 613-626.
11. Cressman, G.P, 1959. An operational objective analysis system. Mon. Wea. Rev., 87(10), 367-374.
12. Daoud, A.B., Sauquet, E., Bontron, G., Obled, C., Lang, M., 2016. Daily quantitative precipitation forecasts based on the analogue method: Improvements and application to a French large river basin. Atmospheric Research, 169, 147-159.
13. Delle Monache, L., Eckel, F.A., Rife, D.L., Nagarajan, B., Searight, K., 2013. Probabilistic weather prediction with an analog ensemble. Monthly Weather Review, 141(10), 3498-3516.
14. Delle Monache, L., Nipen, T., Liu, Y., Roux, G., Stull, R., 2011. Kalman filter and analog schemes to post-process numerical weather predictions. Mon. Wea. Rev., 139, 3554–3570.
15. FAO, 2015. FAOSTAT database. Food and Agriculture Organization of the United Nations, Statistics Division, available at http://faostat3.fao.org/home/E.
16. Ferreira, J., Pardini, R., Metzger, J.P., Fonseca, C.R., Pompeu, P.S., Sparovek, G., Louzada, J., 2012. Towards environmentally sustainable agriculture in Brazil: challenges and opportunities for applied ecological research. Journal of Applied Ecology, 49(3), 535-541.

17. Fu, R., Yin, L., Li, W., Arias, P.A., Dickinson, R.E., Huang, L., Chakraborty, S., Fernandes, K., Liebmann, B., Fisher, R., Myneni, R.B., 2013. Increased dry-season length over southern Amazonia in recent decades and its implication for future climate projection. Proceedings of the National Academy of Sciences, 110(45), 18110-18115.
18. Glahn, H.R., 1985. Yes, precipitation forecasts have improved. Bulletin of the American Meteorological Society, 820-830.
19. Glahn, H.R., Lowry, D.A., 1972. The use of model output statistics (MOS) in objective weather forecasting. J. Appl. Meteorol., 11(8): 1203-1211.
20. Gneiting, T., 2014. Calibration of medium-range weather forecasts. European Centre for Medium-Range Weather Forecasts.
21. Gneiting, T., Balabdaoui., F., Raftery, A.E., 2007. Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69(2), 243-268.
22. Grimm, A.M., Zilli, M.T., 2009. Interannual variability and seasonal evolution of summer monsoon rainfall in South America. J. Climate, 22(9), 2257-2275.
23. Hagedorn, R., Buizza, R., Hamill, T.M., Leutbecher, M., Palmer, T.N., 2012. Comparing TIGGE multimodel forecasts with reforecast - calibrated ECMWF ensemble forecasts. Quarterly Journal of the Royal Meteorological Society, 138(668), 1814-1827.
24. Hagedorn, R., Hamill, T.M., Whitaker, J.S., 2008. Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-meter temperatures. Monthly Weather Review, 136(7), 2608-2619.
25. Hamill, T.M., 2012. Verification of TIGGE multimodel and ECMWF reforecast-calibrated probabilistic precipitation forecasts over the contiguous United States. Monthly Weather Review, 140(7), 2232-2252.
26. Hamill, T.M., Hagedorn, R., Whitaker, J.S., 2008. Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. Monthly weather review, 136(7), 2620-2632.
27. Hamill, T.M., Whitaker, J.S., 2006. Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. Mon Weather Rev, 134(11): 3209-3229.

28. Hamill, T.M., Bates, G.T., Whitaker, J.S., Murray, D.R., Fiorino, M., Galarneau Jr, T.J., Zhu, Y., Lapenta, W., 2013. NOAA's second-generation global medium-range ensemble reforecast dataset. Bulletin of the American Meteorological Society, 94(10), 1553-1565.
29. Hamill, T.M., Scheuerer, M., Bates, G.T., 2015. Analog probabilistic precipitation forecasts using GEFS reforecasts and climatology-calibrated precipitation analyses. Monthly Weather Review, 143(8), 3300-3309.
30. Hamill, T.M., Whitaker, J.S., Mullen, S.L., 2006. Reforecasts - An important dataset for improving weather predictions. B. Am. Meteorol. Soc., 87(1): 33-+.
31. Hamilton, S.K., 2002. Hydrological controls of ecological structure and function in the Pantanal wetland (Brazil). The Ecohydrology of South American Rivers and Wetlands. International Association of Hydrological Sciences, Special Publication, 6, 133-158.
32. Huffman, G.J., Adler, R.F., Morrissey, M.M., Bolvin, D.T., Curtis, S., Joyce, R., McGavock, B., Susskind, J., 2001. Global precipitation at one-degree daily resolution from multisatellite observations. Journal of hydrometeorology, 2(1), 36-50.
33. Janowiak, J.E., Bauer, P., Wang, W., Arkin, P.A., Gottschalck, J., 2010. An evaluation of precipitation forecasts from operational models and reanalyses including precipitation variations associated with MJO activity. Monthly Weather Review, 138(12), 4542-4560.
34. Jones, C., Schemm, J.K.E., 2000. The influence of intraseasonal variations on medium-to extended-range weather forecasts over South America. Monthly Weather Review, 128(2), 486-494.
35. Leal, II.R., da Silva, J., Cardoso, M., Tabarelli, M., Lacher, T.E., 2005. Changing the course of biodiversity conservation in the Caatinga of northeastern Brazil. Conservation Biology, 19(3), 701-706.
36. Liebmann, B., Allured, D., 2005. Daily precipitation grids for South America. Bull. Amer. Meteor. Soc., 86, 1567-1570.
37. Liebmann, B., Allured, D., 2006. Reply. Bulletin of the American Meteorological Society, 87(8), 1096-1096.
38. Marengo, J.A., Liebmann, B., Kousky, V.E., Filizola, N.P., Wainer, II.C., 2001. Onset and end of the rainy season in the Brazilian Amazon Basin. J. Climate, 14(5), 833-852.

39. Medina, H., Tian, D., Srivastava, P., Pelosi, A., Chirico, G.B., 2018. Medium-range reference evapotranspiration forecasts for the contiguous United States based on multi-model numerical weather predictions. Journal of Hydrology, 562, 502-517.
40. Mohr, K.I., Zipser, E.J., 1996. Mesoscale convective systems defined by their 85-GHz ice scattering signature: Size and intensity comparison over tropical oceans and continents. Monthly weather review, 124(11), 2417-2437.
41. Moura, A.D., Shukla, J., 1981. On the dynamics of droughts in northeast Brazil: Observations, theory and numerical experiments with a general circulation model. Journal of the atmospheric sciences, 38(12), 2653-2675.
42. Murphy, A.H., 1973. Hedging and skill scores for probability forecasts. Journal of Applied Meteorology, 12(1), 215-223.
43. Pelosi, A., Medina, H., van den Bergh, J., Vannitsem, S., Chirico, G.B., 2017. Adaptive Kalman filtering for post-processing ensemble numerical weather predictions. Mon Weather Rev, doi.org/10.1175/MWR-D-17-0084.1
44. Rao, V.B., Hada, K., 1990. Characteristics of rainfall over Brazil: Annual variations and connections with the Southern Oscillation. Theoretical and applied climatology, 42(2), 81-91.
45. Ratter, J.A., Ribeiro, J.F., Bridgewater, S., 1997. The Brazilian cerrado vegetation and threats to its biodiversity. Annals of Botany, 80(3), 223-230.
46. Roesch, L.F.W., Vieira, F.C.B., Pereira, V.A., Schünemann, A.L., Teixeira, II.F., Senna, A.J.T., Stefenon, V.M., 2009. The Brazilian Pampa: a fragile biome. Diversity, 1(2), 182-198.
47. Ruiz, J., Saulo, C., Kalnay, E., 2009. Comparison of methods used to generate probabilistic quantitative precipitation forecasts over South America. Weather and Forecasting, 24(1), 319-336.
48. Silva, V.B., Kousky, V.E., Shi., W., Higgins, W., 2007. An improved gridded historical daily precipitation analysis for Brazil. Journal of Hydrometeorology, 8(4), 847-861.
49. Su, X., Yuan, H., Zhu, Y., Luo, Y., Wang, Y., 2014. Evaluation of TIGGE ensemble predictions of Northern Hemisphere summer precipitation during 2008–2012, J. Geophys. Res. Atmos., 119,7292–7310, doi:10.1002/2014JD021733.
50. Subramanian, A., Weisheimer, A., Palmer, T., Vitart, F., Bechtold, P., 2017. Impact of stochastic physics on tropical precipitation in the coupled ECMWF model. Quarterly Journal of the Royal Meteorological Society, 143(703), 852-865.

51. Tian, D., Martinez, C.J., 2012a. Comparison of two analog-based downscaling methods for regional reference evapotranspiration forecasts. J. Hydrol., 475: 350-364.
52. Tian, D., Martinez, C.J., 2014. The GEFS-based daily reference evapotranspiration (ET_0) forecast and its implication for water management in the southeastern United States. J. Hydrometeorol., 15(3): 1152-1165.
53. Voisin, N., Schaake, J.C., Lettenmaier, D.P., 2010. Calibration and downscaling methods for quantitative ensemble precipitation forecasts. Weather and Forecasting, 25(6), 1603-1627.
54. Wetterhall, H., Bao, D.P., Cloke, H., Li, Z.J., Pappenberger, F., Hu, Y.Z., Manful, D., Huang, Y.C., 2010. Ensemble forecasting using TIGGE for the July-September 2008 floods in the Upper Huai catchment: A case study. Atmos. Sci. Lett., 11(2), 132–138, doi:10.1002/asl.270.
55. WHO, World Health Organization, 2014. A global brief on vector-borne diseases.
56. Wilks, D.S., Hamill, T.M., 2007. Comparison of ensemble-MOS methods using GFS reforecasts. Monthly Weather Review, 135(6), 2379-2390.
57. Wilks, D.S., 2006. Comparison of ensemble-MOS methods in the Lorenz'96 setting. Meteorological Applications, 13(3), 243-25.
58. Wilks, D.S., 2006. Statistical Methods in Atmospheric Sciences, 3rd ed. International geophysics series; v. 100, Elsevier, Amsterdam.
59. Xavier, A.C., King, C.V., Scanlon, B.R., 2016. Daily gridded meteorological variables in Brazil (1980–2013). International Journal of Climatology, 36(6), 2644-2659.
60. Zhou, Y., Hejazi, M., Smith, S., Edmonds, J., Li, H., Clarke, L., Calvin, K., Thomson, A., 2015. A comprehensive view of global potential for hydro-generated electricity. Energy & Environmental Science, 8(9), 2622-2633.

CHAPTER V: AN OPTIMIZED MODIS-BASED FRAMEWORK FOR IN-SEASON COUNTY- AND STATE-LEVEL CORN YIELD FORECASTING IN THE U.S. CORN BELT

Abstract: Accurate and timely prediction of corn yields over the Corn Belt regions of the United States is important for decision making regarding food and energy marketing strategies and management of shortages. While statistical models based on moderate resolution imaging spectroradiometer (MODIS) data sets have been commonly used to address this, improved forecasting frameworks enabling to more effectively handling these massive, yet redundant, datasets are needed. This work aimed to develop an optimized framework for the MODIS-based mid-season (2-2.5 months in advance) corn yield forecasting over five major producers states of the United States: Illinois, Indiana, Iowa, Nebraska and Ohio. To achieve this goal, we developed and evaluated the county- and state-level corn yield forecasting considering multiple MODIS products, machine learning techniques, model domains, product subsets, and temporal resolution of pixel composites. The results showed that the performance of the state-level forecasts was often better than the county level forecast. The elastic net and random forest models with multi-temporal EVI composites did not outperform simple linear regression models based on the single latest EVI composite in mid-season. The model domains (i.e. the entities upon which the model is calibrated) that worked best at the county-level performed often suboptimally when aggregating the forecasts, with the choice of the domain particularly affecting the forecasting performance in Nebraska. The performance was instead practically insensitive to the temporal resolutions tested (1-day and 16-days). Compared to the EVI-based forecasts, the NDVI based forecasts performed worse in Illinois, Indiana and Ohio, better in Iowa and similar in Nebraska, while the LAI and FPAR-based forecasts performed poorly over most regions. The mean annual percent errors of the best forecasting framework were between 3% and 5%, which were lower than the mid-season National Agricultural Statistical Service (NASS) forecasts for any of the states.

1 INTRODUCTION

Corn is the most produced cereal worldwide, and one of the most important crops for humanity. It is an essential component of the diets of humans and animals, and can be refined into several bio-products including ethanol, high-fructose corn syrup or even bio-based plastics. The United States is largely the main corn producer providing about 27% of global production and a major player in the world corn trade market, with between 10 and 20 percent of its corn production exported to other countries (FAO, 2019). While corn is grown in most states in the U.S., its production is mostly concentrated in the Corn Belt, with Iowa, Illinois and Nebraska together providing over 43% of the total production in the U.S. Changes in corn productivity over the Corn Belt have dire domestic and worldwide implications. Therefore, accurate forecasting of corn productivity within the growing season over the Corn Belt provides important information to improve food accessibility risk management, which plays a key role in global markets, policy and decision-making.

A common crop yield forecast strategy is to use statistical models based on large datasets of remotely sensed canopy spectral data (Horie et al., 1992). Progress in the remote sensing infrastructure has led to extraordinary advances in mapping and monitoring in numerous agriculture-related activities, including crop yield forecasting. Datasets from the Moderate-Resolution Imaging Spectroradiometer (MODIS) program have been widely used for these applications (e.g., Funk and Budde; Mkhabela et al., 2011; Kogan et al., 2013; Son et al., 2013; Jaafar, and Ahmad, 2015). MODIS has been providing medium-resolution, multi-spectral, daily coverage imagery for about two decades, which makes it especially attractive for the crop yield forecasting over large areas. Unlike high-resolution products such as Landsat or Sentinel-2, MODIS operationally produces temporally aggregated images (Didan et al., 2015), termed as composites, which consider the most reliable observation within a time window and therefore are little affected by cloud cover. Moreover, MODIS not only delivers vegetation indices such as normalized difference vegetation index (NDVI) or the Enhanced vegetation index (EVI), which have proven to be useful in predicting yields, but also biophysical parameters such as Leaf Area Index (LAI) and the fraction of absorbed photosynthetic active radiation (FPAR), which also have a close relationship with productivity (Rembold et al., 2013).

The development of MODIS-based crop yield-forecasting frameworks is nonetheless challenging. It leads to deal with multiple issues that may cause model discrepancies, which have been insufficiently addressed in the literature. Collinearity (redundancy) within multi-temporal composites of any MODIS product is an important issue to consider when setting the crop yield-forecasting framework. Numerous applications avoid collinearity issues by considering a unique timing composite member (e.g. Fun and Budde, 2009; Kogan et al. 2013; Bolton and Friedl, 2013) or a weighted average of few (2-3) members (Hochheim and Barber, 1998; Mkhabela et al., 2005, 2011; Lobus et al., 2002), mostly from the peak of the growing season, on simple regression. However, this strategy might be suboptimal compared with others based on techniques that can explicitly deal with collinearity issues, such as elastic net (Zou et al., 2005) and random forest (Breiman, 2001). Moreover, it is unclear if subsettings considering longer time series of MODIS composites as predictors are advantageous compared with traditional subsettings based on short time series. Crop yields reflect an aggregated response to multiple environmental and management factors throughout a whole season. Studies (e.g. Johnson 2014) show that corn yields over the Corn Belt correlate well with vegetation indices from early stages. Therefore, more research is needed to evaluate how different regression techniques and different product subsets affect forecast performance.

Redundancy between subsets of different MODIS products can be also high. While a large amount of MODIS products provides considerable flexibility for modelers, the choice of the optimal dataset becomes difficult. Studies (Bolton and Friedl, 2013; Johnson, 2016) suggest that EVI better predicts corn yield over the Corn Belt than NDVI, LAI, and FPAR, but is unclear if it can be generalized for every state, i.e., if different states respond the same way to these products. Since states may account for different environmental conditions, management strategies, and policies (e.g. Singer et al., 2007; Tannura et al., 2008), there is likely to be more than one best product. Considering the impact that corn productivity on any of the states of the Corn Belt has on food and energy availability and prices, more comprehensive, state-specific analyses, are needed to compare the performance of crop yield forecasting frameworks based on different MODIS products such as NDVI, EVI, LAI, and FPAR.

Properly setting the model domain of statistical models is also challenging. Statistical models can be built locally, i.e., based on a time-series data over a single unit, or regionally, in which the model is built based on information at different locations and eventually different times.

The more confined the training domain, the more able the model is to seize the behavior over the particular location, but the shorter and probably the less representative the dataset for training. A usual strategy in crop yield forecasting based on remote sensing data is to produce a common model for a region sharing similar characteristics. For example, Bolton and Friedl (2013) grouped the counties as semi-arid or non-semi-arid and produced a model for each group independently. Mkhabela et al. (2011) grouped the units based on soil type into three agro-climatic zones: sub-humid, semi-arid and arid, while Johnson et al. (2016) used a clustering strategy for grouping the forecast units based on their crop yield time series. However, a systematic evaluation of the impact of model domains on yield forecast has not been considered in previous studies. For example, while users may be interested in the forecast performance at different scales, an important but poorly addressed question is if a model domain, which is optimum at the short scale, remains optimum when aggregating the forecasts to a larger scale. Since the performance of statistical models may change with explanatory variable and spatial scale (Lobell and Burke, 2010), a systematic analysis of how the model domains affect the crop yield forecasting over the Corn Belt can be of great help.

In addition to individual daily scenes, MODIS provides consolidated 8-day and 16-day composites images. The 16-day-composites may be considerably less noisy than the 8-days composites (and logically the daily scenes), as the 8-day compositing period is sometimes shortened to get a clear sky value for every pixel (e.g., Sakamoto et al., 2013). However, the improvements in noise reduction through the 16-day compositing come at a cost in temporal resolution. Johnson et al. (2016) reported that the choice of 8-days or 16-days composites of NDVI had minor implications on the crop yield forecasting of several crops, including corn. However, Guindin-Garcia (2012) showed that the use of untreated 16-days composites of NDVI and EVI lead to inaccuracies of estimated corn leaf area index over the Corn Belt, compared with intermediate products that consider the true day of the pixel composites. Since MODIS 16-day composites of EVI and NDVI include a layer with the true day of every pixel, it is easy to obtain interpolated daily assessments of the product values that may help to circumvent the issues with temporal resolution of the untreated composite. While 16-day product composites are useful for crop yield forecasting (e.g. Kogan et al., 2013), the implications of their low temporal resolution on forecasting accuracy are poorly understood. A question is if improvements are feasible by considering this data layer in the forecasting framework compared to the frameworks that ignore

it. Therefore, the comparison between the forecast performance considering the untreated coarse resolution (16-day) MODIS products and the interpolated high-resolution datasets may help to better understand the shortcomings because of the loss in temporal resolution of MODIS 16-day composites, and the benefits users may expect by considering the true day of the composite in the forecasting framework.

This work aims to construct an optimized framework for MODIS-based corn yield forecasts over major producer states of the U.S., by considering multiple machine learning techniques, product subsets, model domains, and temporal resolutions. It is also aimed at evaluating and comparing the performance of the optimized framework based on the MODIS NDVI, LAI, and FPAR products. To the authors' knowledge, this is the first study comprehensively evaluating the state-specific impacts of these factors on satellite-based, large-scale corn yield forecasts.

2 METHODS

2.1 Study region and datasets

The study focuses on the in-season forecasting of the county-level corn yields over five states of the U.S.: Nebraska, Iowa, Illinois, Indiana, and Ohio over 16 years, from 2002 to 2017. These five major corn-producing states together provide about 56 percent of U.S. corn production. In this study, we used composites of the normalized difference vegetation index (NDVI) and the enhanced vegetation index (EVI), produced at 16-day intervals with 250 m spatial resolution (Didan, 2015), based on the MODIS MOD13Q1 (TERRA satellite) and MYD13Q1 (AQUA satellite) data products. The NDVI has been designed to standardize vegetation index values to between −1 and +1and is expressed as

$$\mathrm{NDVI} = \frac{\mathrm{NIR}-\mathrm{Red}}{\mathrm{NIR}+\mathrm{Red}} \quad \text{(V.1)}$$

while the EVI is commonly expressed as

$$\mathrm{EVI} = 2.5\frac{\mathrm{NIR}-\mathrm{Red}}{\mathrm{NIR}+6\mathrm{Red}-7.5\times\mathrm{Blue}+1} \quad \text{(V.2)}$$

where the NIR and Red and Blues are atmospheric-corrected surface reflectances at the specific bands. The EVI is usually less affected by the atmospheric effects than the NDVI, as it accounts for the difference in blue and red reflectances. In addition to the EVI and NDVI, we also

retrieved the layer with the composite day for every pixel, which is also available in the MOD13Q1 and MYD13Q1 products. This layer indicates the date of acquisition of the reflectances used in vegetation indices computation.

We also used the Leaf Area Index (LAI) and the fraction of the photosynthetic active radiation (FPAR) generated on 8-day intervals at 500m resolution (Myneni et al. 2016) based on the MOD15A2H products. LAI is defined as the one-sided green leaf area per unit ground area in broadleaf canopies and represents a common measure of the crop phenology. FPAR is the fraction of photosynthetically active radiation (400-700 nm) absorbed by green vegetation. Both LAI and FPAR can be seen as subproducts of NDVI (Myneni et al., 2016) that parameterize the quantity and quality of the canopy cover, similarly as the vegetation indices. For details about the methodology used for retrieving LAI and FPAR, see Myneni et al. (2016).

2.2 Forecasting framework

To create within season forecast, we considered the composites available between early-season to mid-season, specifically from the beginning of June (day of the year, DOY 153) to the end of July (DOY 209), involving eight composites. The time series of EVI and composite days also considered the composite at DOY 145, which was used for modeling the crop phenology (see below). The DOY indicates the mean day over 16-day or 8-day time intervals. We generated county aggregates of every MODIS product, by computing the mean of the product values over the areas with corn, which were identified using Cropland Data Layer (CDL), the land cover data layer, hosted on CropScape (https://nassgeodata.gmu.edu/CropScape/). We used the "pyModis" (Delucchi and Neteler, 2013) and "GDAL" packages (GDAL/OGR contributors, 2018), in Python, for the manipulation of the MODIS data. We used final yields reported by the USDA-National Agricultural Statistics Service (NASS) as observed yields. We also use the NASS planted acreage information released in June, for aggregating the county-level forecasts to state-level, and the state level forecasting reports released in August, for comparison purposes. The NASS forecasts reflect the condition at the beginning of that month (Vogel and Bange, 1999) and are based on assessments of planted and harvested area and two types of yield estimates, a farmer-reported survey, and independent measurements.

We produce in-season forecasts for each county, each state, and each year, and evaluated them against the observed yields. The state forecasts were produced by aggregating the county

forecasts into state averages. To avoid overfitting while testing the forecasting models in an operational mode, we performed leave-one-year-out cross-validation by holding one year of data out for model testing and training the models using all the other years of data. In general, a yield forecast ($\hat{y}_{i,j}$) at the i^{th} $(i = 1 \dots N)$ county and $j^{th}(j = 1, \dots, M)$ year is produced as

$$\hat{y}_{i,j} = \hat{f}(x_{i,j}) + \hat{\tau}_{i,j} \quad \text{(V.3)}$$

with $x_{i,j}$ describing a p-dimensional feature space and $\hat{\tau}_{i,j}$ is a trend term accounting for the technological improvements across years and counties. The $\hat{\tau}_{i,j}$ term was estimated independently based on the linear dependence of a 28-year (from 1990 to 2017) data series of yields on time. Figure V.1 shows a map with the slopes of those relationships. $\hat{f}$ represents a regression function which is characteristic of every machine learning technique whose parameters (in a wide sense) are obtained through a numerical optimization via minimization of a loss function $L(f)$ with the generic form:

$$L(f) = \sum_{\upsilon=1}^{\Upsilon_\xi} \sum_{\xi=1}^{\Xi} L\left(y_{\xi,\upsilon} - f(x_{\xi,\upsilon})\right) \quad \text{(V.4)}$$

considering the training observations over a set of Ξ counties, and Υ_ξ years (where the subscript is to indicate that different counties can be represented in different subsets of years).

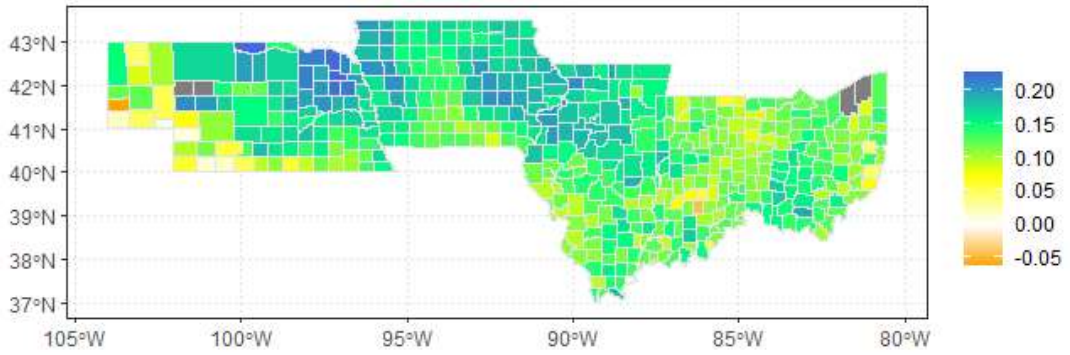

Figure V.1. Map of the slope (in Mg ha^{-1} y^{-1}) of the linear relationship between the NASS county yields and the years based on the 1990-2017 time series. The information for the counties in gray is missing.

Figure V.2 shows a schematic framework for finding the optimized forecasting scheme by comparing different datasets, models, and configurations for in-season corn yield forecasting. The first part of the framework is to evaluate the impact of different factors (including resolution of the composite period, product subsetting, machine learning technique, and model domains) on EVI-based corn yield forecasting. We started from testing the EVI-based corn yield forecasting because

previous studies suggested that EVI served as a better predictor than NDVI, LAI, and FPAR (Bolton and Friedl 2013, Johnson 2016).

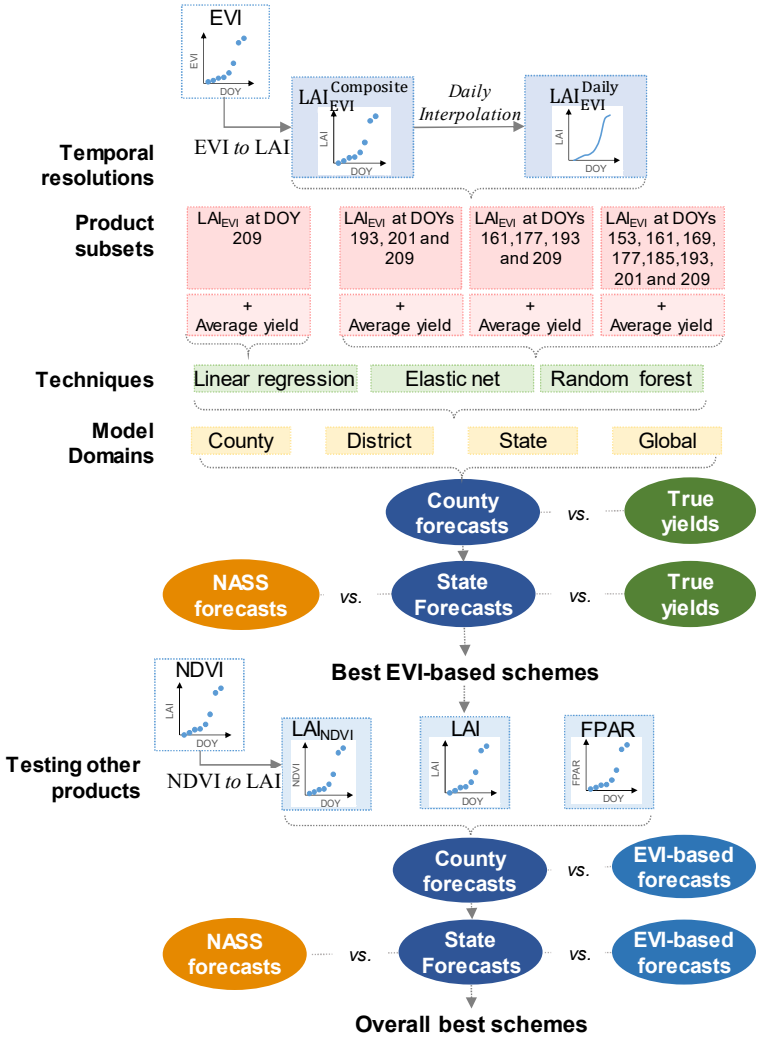

Figure V.2. Schematic diagram of the proposed frameworks for in-season corn yield forecasting

We transformed the county level EVI into estimates of the leaf area index (LAI_{EVI}) using the methods employed by Guindin-Garcia et al. (2012):

$$LAI_{EVI} = -1.84 + 9.05 \times EVI + 0.94 \times EVI^2 \quad (V.5)$$

Guindin-Garcia et al. (2012) found a strong linear relationship between LAI estimated with Eq. 5 and the corn yields, based on field experiments over a region of the Corn Belt. It makes LAI_{EVI} subproducts more suited as predictors in linear regression models than EVI. Therefore, subsets of LAI_{EVI} are used as our actual predictors for yield. Notice that we use LAI_{EVI} (or LAI_{NDVI},

see section 2.2.5) to referring to the LAI based on the EVI (or NDVI), while we use only "LAI" when referring to the operational MODIS LAI.

In the second part of the framework, we replace EVI in the best EVI-based scheme by NDVI, LAI, and FPAR and compare the forecast performance with the different products. In the sections below, we provide details about the implementation of the experiments for testing those factors.

2.2.1. Temporal resolutions of the LAI_{EVI} series

In addition to the original time series of LAI_{EVI} considering a 16-day temporal resolution (the coarse resolution), we generated a time series of LAI_{EVI} daily (the high resolution) that consider the actual acquisition days of the MODIS image. We made this by fitting the dependence of the original LAI_{EVI} on the acquisition days using a logistic function (as Zhang et al., 2003) for every county with the constrain that the minimum LAI_{EVI} be zero, and producing interpolated daily estimates of the LAI_{EVI}. The original LAI_{EVI} at DOY 145 participated in the fitting of the logistic function but it was excluded from the subsets of forecasting features.

2.2.2. Subsetting strategies

We tested the model response (in Eq. 3) to different time series configurations of the LAI_{EVI} at both temporal resolutions. The subsets considered:

A single vector of LAI_{EVI} based on the last composite considering either the mean composite day ($\langle\overline{209}\rangle$) or the actual composite day ($\langle 209\rangle$).

A three-member subset of LAI_{EVI} considering the last three available composites (either $\langle\overline{193}, \overline{201}, \overline{209}\rangle$ or $\langle 193,201,209\rangle$)

A four-member time series of LAI_{EVI} based on the AQUA products between mean DOY 161 and 209 ($\langle\overline{161}, \overline{177}, \overline{193}, \overline{209}\rangle$). In this case, we only considered the LAI_{EVI} estimates based on the mean composite days.

An eight-member time series accounting for both the TERRA and AQUA products between mean DOY 153 and 209 (either $\langle\overline{153}\text{ to }\overline{209}\text{, in steps of 8}\rangle$ or $\langle 153\text{ to }209\text{, in steps of 8}\rangle$).

In addition to the subsets of $LAI_{EVI,}$ the array included the average yield as another predictor, except for the county-based domain (see section 2.2.4). We computed the average corn yield for every county by taking the mean over the last least five years, excluding the testing year.

For example, if we wanted to forecast the yields in 2010, the average yield corresponding to a specific county in 2014 considered the yields at that county in 2008, 2009, 2011, 2012 and 2013. Therefore, we had 16 replicas of the subsets of features (as many as the total number of years), which slightly differ in the average yields. The average yields involve useful information about the yield response to a specific ecoregion and are easy to obtain in real-life applications.

2.2.3 Machine learning models

We tested three machine-learning models (Eq. 3): linear regression (LR, e.g. Montgomery et al.; 2012), elastic net (EN, Zou et al., 2005), and random forest (RF, Breiman, 2001), which has proven efficient in a broad range of applications. Linear regression used unique timing composite member of LAI_{EVI} as explanatory variables, while elastic net and random forest considered multi-temporal arrays. In Table V.1, we show the tested configurations.

Table V.1. Configurations of the arrays of predictors, the temporal resolution of the composites and the machine learning techniques.

N	Time series for the two resolutions	Linear model (LM)	Elastic net (EN)	Random Forest (RF)
1	$\langle\overline{209}\rangle$ and $\langle 209\rangle$	✓		
3	$\langle\overline{193}, \overline{201}, \overline{209}\rangle$ and $\langle 193,201,209\rangle$		✓	✓
4	$\langle\overline{161}$ to $\overline{209}$, each 16$\rangle$		✓	✓
8	$\langle\overline{153}$ to $\overline{209}$, each 8$\rangle$ and $\langle 153$ to 209, each 8$\rangle$		✓	✓

For a finite set of, in general correlated, *p* predictors, the EN and RF techniques provide larger reductions in variance than LR, at the cost of slightly larger bias, ultimately reducing the prediction errors. Unlike the RF technique, both LR and EN assume that the underlying association between the predictors and the predictand is linear. We use the 'glmnet' (Freidman et al., 2010) and the 'randomForest' (Liaw and Wiener, 2002) packages in R (R Core Team, 2016) to implement the EN and RF techniques, respectively, and the 'stats' package, which is part of base R, to implement the linear regression technique. Using glmnet we implement a *k*-fold cross-validation based on the training data and selected the regularization term that provided minimum mean cross-validated error. We used the default number of folds (10), except for the county-based models (see the following section), for which we considered four-folds. The optimal regularization term was chosen based on the 'gmlnet' own sequence. The number of trees in random forest was set to 3000, while the other parameters, such as the number of variables sampled as candidates at each split and the size of the sample to draw (Liaw and Wiener, 2002), were set to default values.

2.2.4 Model domains

We tested four model domains: the county-, district-, state-, and global-domains, which differ from each other on the spatial extent within which the observations were chosen for model training. The county-based model is equivalent to a time series model (as in Lobell and Burke, 2010) in which the model is built based on temporal data series from a single unit (county). It entails that the loss function $L(f)$ (Eq. 6) considers the observations over the same i^{th} county for testing and the $M-1$ training years

$$L(f) = \sum_{v=1\dots M, v\neq j} L\left(y_{i,v} - f\left(x_{i,v}\right)\right) \quad \text{(V.6)}$$

Using the district- (state-) based model, $L(f)$ considers the observations over the subset of the counties belonging to the same κ^{th} district (state) Δ_κ as the i^{th} county:

$$L(f) = \sum_{v=1\dots \Upsilon_\xi, v\neq j} \sum_{\xi=1\dots \Xi,\, \xi \& i \in \Delta_\kappa} L\left(y_{\xi,v} - f\left(x_{\xi,v}\right)\right) \quad \text{(V.7)}$$

For the global model, $L(f)$ considers all the observations available for training. In this work we use "regional models" for generically referring to the district-based, state-based, and the global models. When considering the county-based model domain we excluded the average yield as a predictor, since it is uninformative (in principle homogeneous) in that domain.

Finally, we evaluated the performance of the county- and state-level forecasts accounting for multiple subsets of predictors, machine-learning techniques, model domains, and temporal resolutions to identify the best forecasting scheme. The state-level forecasts were produced by considering the weighted mean of county-level forecasts, with the weights being the fraction of the area sown at every county to the total area sown.

2.2.5 Testing other MODIS products

Besides EVI, we also tested other MODIS products for in-season corn yield forecasting. We generated NDVI-, LAI-, and FPAR-based corn yield forecasts at both county- and state-level using the schemes that worked best for EVI-based forecasts. Then, we compared these forecasts for the different products to get an overall best scheme. Similarly, as for the EVI (Eq. 5), we transformed the county level NDVI into estimates of the leaf area index ($\mathrm{LAI_{NDVI}}$) using the expressions of Guindin-Garcia et al. (2012)

$$\mathrm{LAI_{NDVI}} = 1.94 - 10.84 \times \mathrm{NDVI} + 16.53 \times \mathrm{NDVI}^2 \quad \text{(V.8)}$$

and used the $\mathrm{LAI_{NDVI}}$ as predictors in the NDVI-based forecasting framework.

2.3. Evaluation of corn yield forecasting framework

In this study, we used the percent error PE and the mean absolute percentage error (MAPE), and the coefficient of determination, to evaluate the forecast performance. The PE and MAPE provide measures of forecast accuracy, while R^2 represents the variance of the yields explained by the forecasts. Equations for calculating PE and MAPE are given below:

$$\text{MAPE} = \frac{1}{K}\sum_{k=1}^{K} \text{PE}_k = \frac{1}{K}\sum_{k=1}^{K} \left| 100 \frac{y_k - \hat{y}_k}{y_k} \right| \quad \text{(V.10)}$$

$$\text{R}^2 = 1 - \frac{\sum_k (y_k - \hat{y}_k)^2}{\sum_k (y_k - \bar{y})^2} \quad \text{(V.11)}$$

where y_k, $\hat{y}_k$ are observed and forecasted yields at the k^{th} ($k = 1 \ldots K$) unit, respectively, and $\bar{y}$ is the average observed yield over the K units.

We firstly illustrate the results of an exploratory analysis to help understand the levels of predictability and to provide insights about the homogeneity in the crop growth dynamics among years and states. It included the analysis of the annual correlation of the yields with the time series of the MODIS composites over the entire region, as well as with the interpolated daily LAI_{EVI} across states. Then we evaluated the response of county- and state-level forecasts to each factor when fixing the other factors. We used pairwise Wilcoxon tests with Bonferroni correction (Johnson and Wichern, 2002) to test if the response was significant. The Wilcoxon test is a nonparametric alternative to Student's t-test for comparing two samples and is especially useful when the sample size is small and the population is not normally distributed. For the state-level forecasts, we compared the performance of our forecasts with the NASS state-level forecasts issued in August. Finally, we compared the performance of the forecast schemes using the NDVI-, LAI- and FPAR-based county-level forecasts with the forecast schemes using the EVI-based forecasts.

3 RESULTS

3.1 Exploratory analysis

Figure V.3 shows the correlations of corn yields with each of EVI, NDVI, LAI and FPAR composites over the forecasting period.

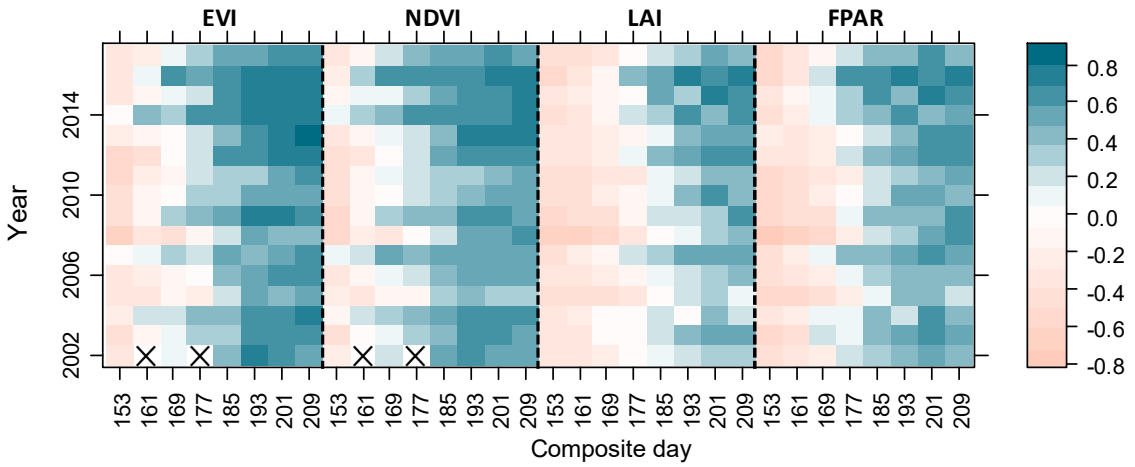

Figure V.3. Correlations between the county EVI, NDVI, LAI and FPAR with the yields for the different composite days and years. The EVI and NDVI at mean DOYs 161 and 177 over 2002 (indicated with "×") were missing.

The correlations with EVI tended to be the largest, closely followed by the correlations with NDVI, while clearly the correlations with MODIS LAI tended to be the smallest. In general, the correlations grow consistently until around DOY 193, while they change little afterward. Similar to Johnson (2014), we found notable negative correlations between the MODIS products and the yields during the early stages of the growing season (Fig. V.3). This is probably because the MODIS vegetation products reflect the climate and management conditions over the regions. For example, northern regions commonly reach higher yields than southern regions because of the better environmental conditions (see Lobell et al., 2014), but they are planted later and therefore exhibit a lower crop development (and then lower EVI, NDVI, etc.) at early stages.

Figure V.4 shows the correlations between the daily LAI_{EVI} (computed with Eq. 3) and the yields across years and states. The maximum correlation tends to be largest in Illinois, probably because the climate variability is also largest so that the changes in yields are more easily resolved by the changes in LAI_{EVI} (i.e., the covariance between LAI_{EVI} and yields is in proportion less affected by random errors). Whereas, the maximum correlations tend to be lowest in Ohio, followed by Indiana, probably because these states have more diversified agriculture, which leads to the composite products are more contaminated by mixed land uses. In general, the day of maximum correlation is highly variable between both the states and the yields. Notice that in Nebraska, where the impact of climate variability on the overall yield variability might be relatively lower because of irrigation, the correlation patterns had some distinct features compared

to other states. From Figures V.3 and V.4 is easy to infer that correlation between composites at late crop growth stages, as well as early stages, is commonly high.

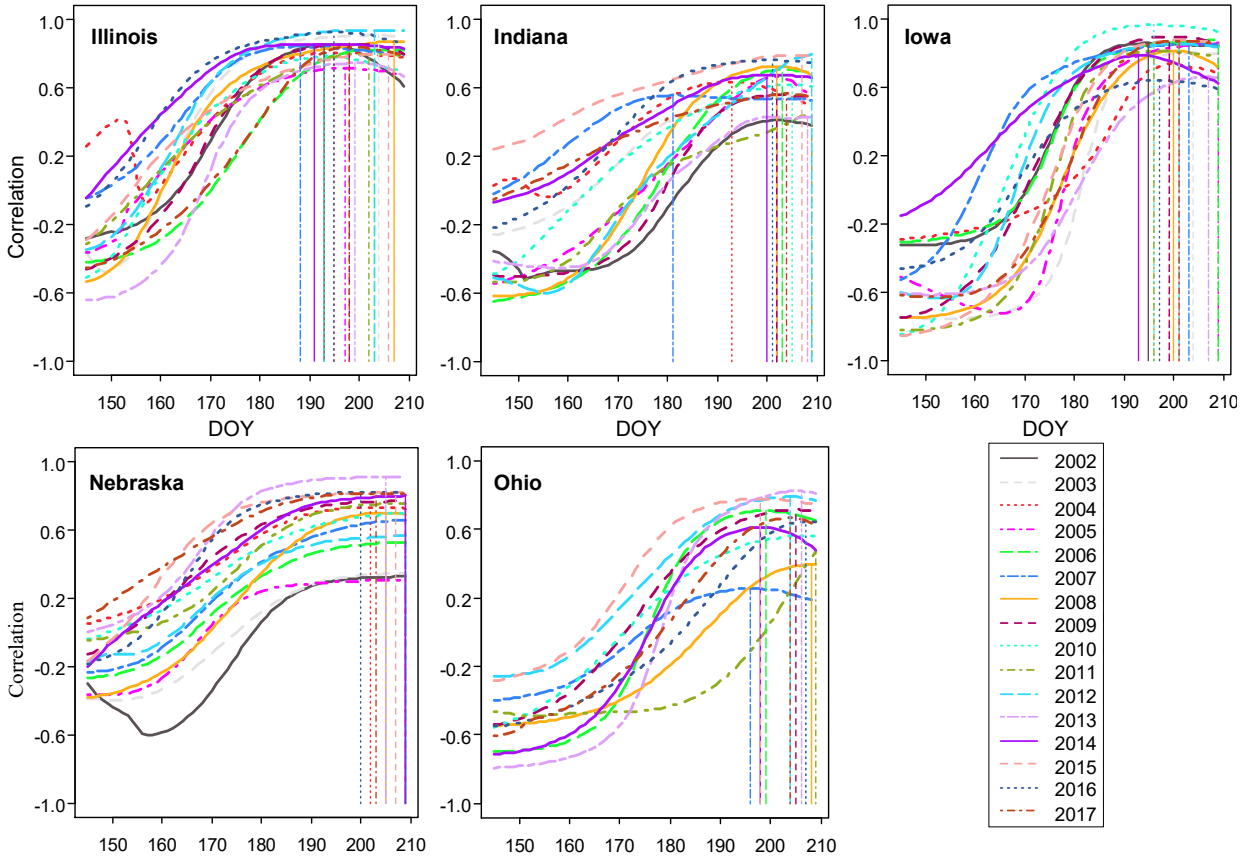

Figure V.4. Correlation between the daily LAI_{EVI} and the yields between DOYs 145 and 209. Vertical lines indicate the DOY of maximum correlations.

Figure V.5 shows the distribution of the yields and LAI_{EVI} based on the last available composite $\langle\overline{209}\rangle$. In general, changes in the mean and variance of the LAI_{EVI} match fairly well with the yields in terms of the mean and variance. In 2012, there was noticeably low crop growth and yields, because of one of the most severe droughts in U.S. history (Mallya et al., 2013). In terms of long-term pattern, the LAI_{EVI} showed a less consistent growing trend across years than the yields, which indicate that the improvements in crop management and the technology have a large influence on annual average yield (e.g. Assefa, 2017), but less impact on the green canopy. In practice, this justified the adopted strategy of considering a trend term in Eq. 3.

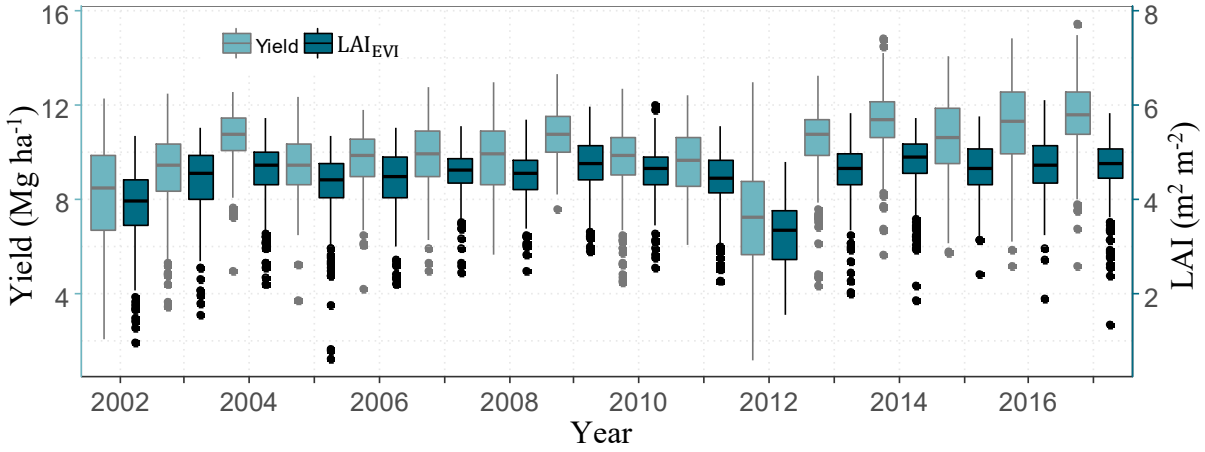

Figure V.5. Distribution of annual yields and the $LAI_{EVI}(\overline{209})$ (see Eq. V.5). The dots represent outliers in the distribution.

3.2 Performance of the EVI-based county-level forecasts

3.2.1 Influence of the spatial domains

Figure V.6 compares the performance county-level forecasts with the different model domains for linear regressions based on $\langle\overline{209}\rangle$ and elastic net and random forest regressions based on $\langle\overline{193}, \overline{201}, \overline{209}\rangle$. The Supplemental Material section shows the performance considering the other tested schemes for linear regression (Fig. V.A1), elastic net (Fig. V.A2) and random forest (Fig. V.A3), respectively. In general, the forecast error was especially variable across years, while the R^2 was variable across both years and states. However, the performance was non-significantly affected by the domains, except when considering the MAPE in Nebraska and the R^2 based on elastic net. The elastic net technique performed poorly at the county domain, most likely due to issues related to the selection of the best regularization term using a small training sample (see Section 2.3.1). The linear regression and the random forest techniques were less sensitive to the choices of model domains than elastic net (see Fig. V.6 and Figs. V.A1-V.A3). The impact of the model domain on the forecasts represented in Figure V.6 was similar to the impact considering the other schemes tested (Figures V.A1- V.A3), suggesting that there was low interaction of the domain with the subset of predictors and the temporal resolution.

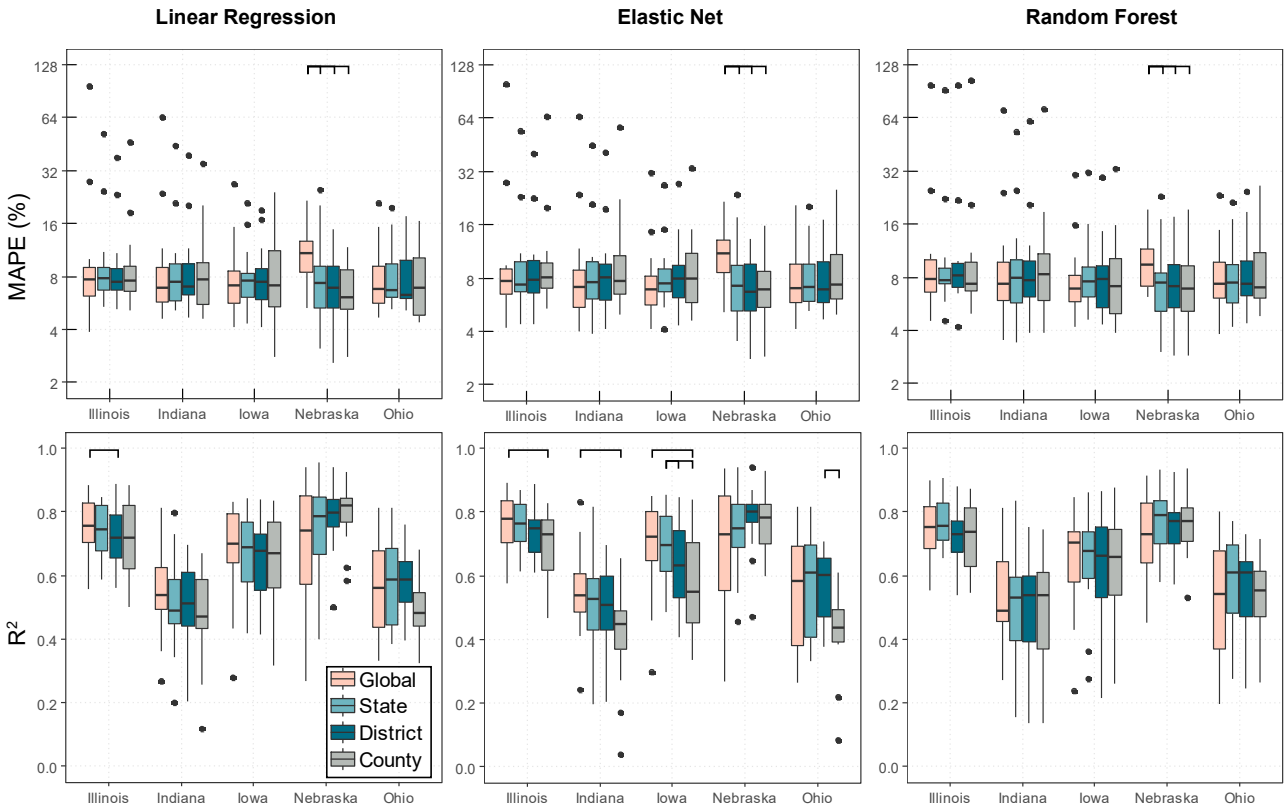

Figure V.6. Distribution of the annual mean absolute percentage error (MAPE) and annual R^2 of the county-level forecasts considering the different spatial domains with the linear regression based on $LAI_{EVI}\langle\overline{209}\rangle$ and the elastic net and the random forest based on $LAI_{EVI}\langle\overline{193}, \overline{201}, \overline{209}\rangle$. The bars in the top of the plots denote significant differences with respect to the domain aligned with the mark on the left, resulting from a pairwise Wilcoxon test analysis with Bonferroni adjustments. The dots represent outliers in the distribution.

Figure V.7 shows the distribution of the annual percent errors and the R^2 for the state-level crop yield forecasts as well as the NASS forecasts. The results of the aggregation for the schemes represented in Figures V.A1-V.A3 are shown in Figures. V.B1-V.B3, also in the Supplemental Material section. In general, the performance at the state level was better than at the county level. Similarly as for the county level, the error was non-significantly different among domains at the state level, except for Nebraska. Notice that the performance considering the domains and machine techniques tested tended to be slightly better than the performance considering the NASS predictions, except for Iowa. The median state-level errors using our forecasts varied approximately between 3 % in Ohio and Nebraska and 5% in Iowa, while for the NASS forecasts it varied between 4% in Ohio and Iowa and 6% for Illinois. The EVI-based forecasts were especially successful in Nebraska, as the percent errors were below 5 percent for most years.

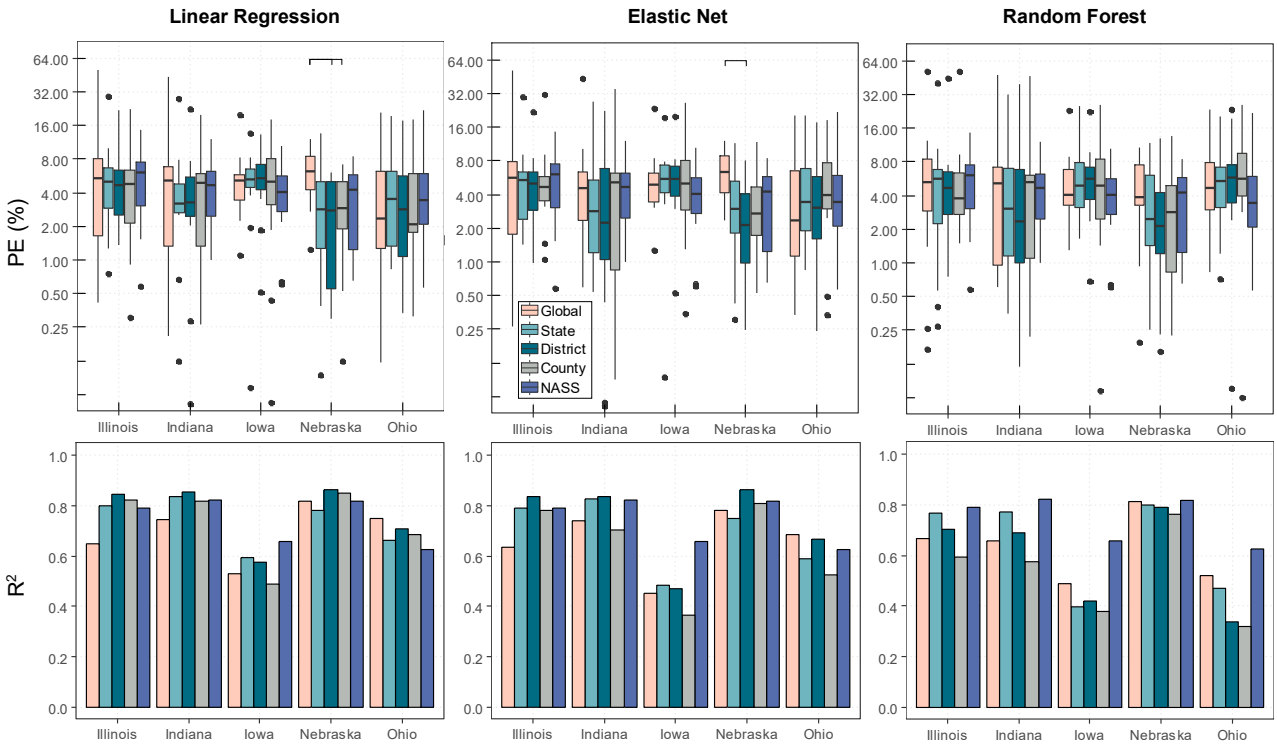

Figure V.7. Distribution of percent error (PE) and R^2 of the annual **state-level** forecasts considering the linear regression based on $LAI_{EVI}\langle\overline{209}\rangle$ and the elastic net and random forests based on $LAI_{EVI}\langle\overline{193}, \overline{201}, \overline{209}\rangle$ for the four spatial domains as well as the percent considering the NASS forecasts. The bars in the top of the plots denote significant differences with respect to the domain aligned with the mark on the left. The dots represent outliers in the distribution.

In Table V.2 we ranked the performance provided by the different model domains from linear regressions based on $\langle\overline{209}\rangle$. At the county level, the global domain was the overall best, followed by the county-based domain, while at the state level the district-based domain was best, followed by the state-based domain. This indicates that the response to the domains is scale-dependent. See that in Nebraska the local domains performed particularly well, while the global performed poorly, which is most likely associated with the differences in water regimes within the states and with other states. When using elastic net and random forest, we observed similar patterns as linear regression, especially with the former technique (data is not shown).

Table V.2. Number of years the different spatial domains provided the best (N_b), the second-best (N_s) and worst (N_w) performance based on the linear regression on $\langle\overline{209}\rangle$. On the right, we weighted the performance for every domain through the expression $\sum_i\left(2 \times N_b^i + N_s^i - N_w^i\right)$, where *i* accounts for the performance metric (MAPE and R^2 for the county level and PE for the state level).

			Best performance				2nd best performance				Worst performance				Weighted performance				Overall Best
			Gl	St	**Dt**	**Ct**	**Gl**	**St**	**Dt**	**Ct**	**Gl**	**St**	**Dt**	**Ct**	**Gl**	**St**	**Dt**	**Ct**	
County-level	MAPE	Illinois	6	3	2	5	1	3	7	5	6	5	1	4	12	3	7	10	Global
		Indiana	7	2	3	4	4	4	5	3	4	4	3	5	31	-2	-3	6	Global
		Iowa	7	1	4	4	3	8	4	1	3	2	4	7	17	4	2	9	Global
		Nebraska	1	4	5	6	0	4	8	4	11	3	1	1	-14	14	11	21	County
		Ohio	4	2	1	5	3	2	6	1	3	3	2	4	3	7	15	-1	District
		Total	**25**	**12**	**15**	**24**	**11**	**21**	**30**	**14**	**27**	**17**	**11**	**21**					
	R^2	Illinois	6	4	2	4	6	4	4	2	1	5	3	7					
		Indiana	10	2	0	4	3	5	7	1	0	5	4	7					
		Iowa	5	3	1	7	5	5	4	2	2	5	4	5					
		Nebraska	3	5	2	6	3	4	7	2	8	1	3	4					
		Ohio	1	4	6	1	3	4	2	3	4	0	1	7					
		Total	**25**	**18**	**11**	**22**	**20**	**22**	**24**	**10**	**15**	**16**	**15**	**30**					
State-level	PE	Illinois	6	0	6	4	2	7	4	3	7	2	2	5	7	5	14	6	District
		Indiana	3	4	3	6	4	5	6	1	7	4	0	5	3	9	12	8	District
		Iowa	4	3	3	6	4	6	5	1	3	2	4	7	9	10	7	6	State
		Nebraska	3	4	7	2	2	5	3	6	11	2	1	2	-3	11	16	8	District
		Ohio	5	0	3	4	2	5	3	2	2	5	0	5	10	0	9	5	Global
		Total	**21**	**11**	**22**	**22**	**14**	**28**	**21**	**13**	**30**	**15**	**7**	**24**					

3.2.2 Response to the machine learning techniques and product subsets

Here we compare the schemes accounting for different algorithms and subsetting strategies using the linear regressions on $\langle\overline{209}\rangle$ optimized for the domain (i.e. based on the domains with largest rankings in Table V.2) as a baseline. Figure V.8 compares the schemes when the elastic net and random forest use 16-day composites of LAI_{EVI}. Analog comparisons, but for when they use interpolated daily LAI_{EVI}'s are provided in the Supplemental Material section. The use of the elastic net and random forest techniques with multi-temporal composites provided no gains compared with the use of linear regression based on a single composite. Elastic net was better than random forest (except with the county-based domain) and, when relied on the best subsets and domains, was as effective as linear regression at both county and state levels. This technique worked noticeably better when based on three composites ($\langle\overline{193}, \overline{201}, \overline{209}\rangle$) than when based on four or eight composites.

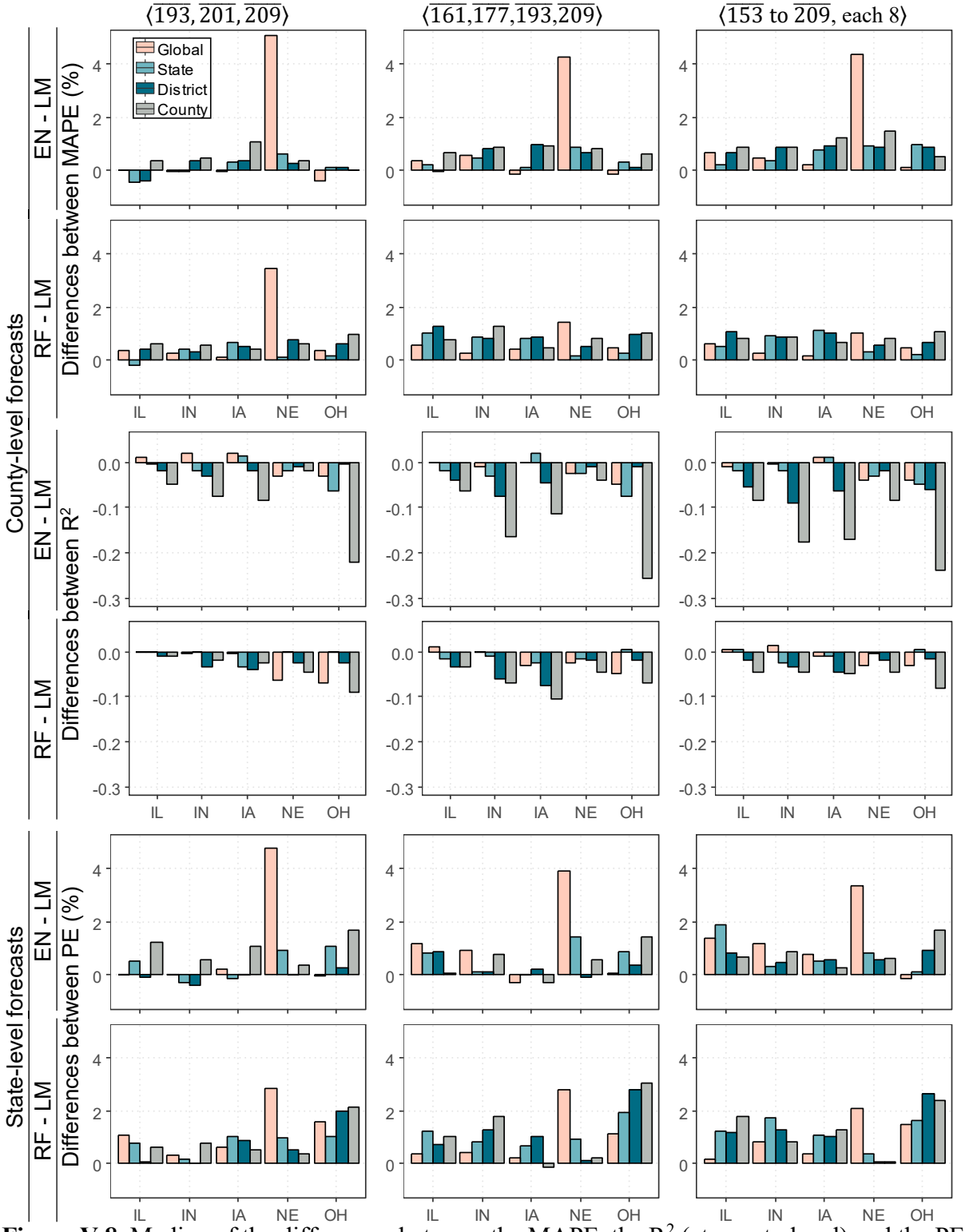

Figure V.8. Median of the differences between the MAPE, the R^2 (at county level) and the PE (at state-level) with elastic net and random forest and the MAPE, the R^2 and the PE with linear regression for the four model domains. The subsets of LAI_{EVI} used by elastic net and random forest are indicated at the top. The linear regressions based on $\langle\overline{209}\rangle$ and the best domains identified in Table V.2.

The random forest technique performed poorly, especially at the state level. For example, the percent errors at the state level in Ohio were on average more than 28 % larger than based on linear regression. This technique also provided lower errors when based on the subsets of three composites than based on four and eight composites, but led to better R^2 considering subsets of eight composites.

3.2.3 Influence of the temporal resolution

In Table V.3 we compare the performance of the county-level forecasts based on composites of different temporal resolution. The differences were generally small and non-significant. For the state level, the impact of the temporal resolution is also negligible (not shown). This suggests that there is a tradeoff between the error product of interpolation of the LAI_{EVI} for the approach based on the actual pixel days and the errors in the temporal resolution with the traditional approach (based on the mean day). This is likely because the logistic approach used for the interpolation tended to over-smooth the variability around the zone of maximum LAI_{EVI} where the curvature is high.

Table V.3. Mean differences between the county-level MAPE and R^2 based on the composites at daily resolution and the MAPE and R^2 based on the coarse temporal resolution for the three techniques (linear regression (LNR), elastic net (EN) and random forest (RF)), and the four spatial domains (global, state, district, and county-based). The "*", "**" and "***" indicates that the differences are significant at α confidence levels 0.95, 0.99 and 0.999 based on a Wilcoxon test.

	MAPE (%)				R^2			
	Global	State	District	County	Global	State	District	County
$\mathrm{LNR}(\langle 209\rangle - \langle\overline{209}\rangle)$	0.03	0.06	0.19	0.15	0.003	0.005	0.008	0.009
$\mathrm{EN}(\langle 193, 201, 209\rangle - \langle\overline{193}, \overline{201}, \overline{209}\rangle)(\%)$	0.04	0.07	0.25	-0.07	-0.002	-0.002	0.002	0.010
$\mathrm{EN}\big(\langle 153 \text{ to } 201, \text{each } 8\rangle - \langle\overline{153} \text{ to } \overline{209}, \text{ each } 8\rangle\big)(\%)$	-0.18	-0.10	-0.14	-0.14	0.005	-0.003	0.016	0.010
$\mathrm{RF}(\langle 193, 201, 209\rangle - \langle\overline{193}, \overline{201}, \overline{209}\rangle)(\%)$	-0.24	-0.08	0.20	0.24	-0.002	-0.009	-0.009	-0.020 ***
$\mathrm{RF}\big(\langle 153 \text{ to } 201, \text{each } 8\rangle - \langle\overline{153} \text{ to } \overline{209}, \text{ each } 8\rangle\big)(\%)$	-0.16	-0.12	0.08	0.04	-0.026 ***	-0.025 ***	-0.018 **	-0.016 **

3.3. Comparing the NDVI, LAI and FPAR forecasts with the EVI based forecasts.

In Figure V.9 we compare the performance of the NDVI-, LAI- and FPAR-based county-level forecasts against the performance of the EVI-based forecasts, based on the linear regression model.

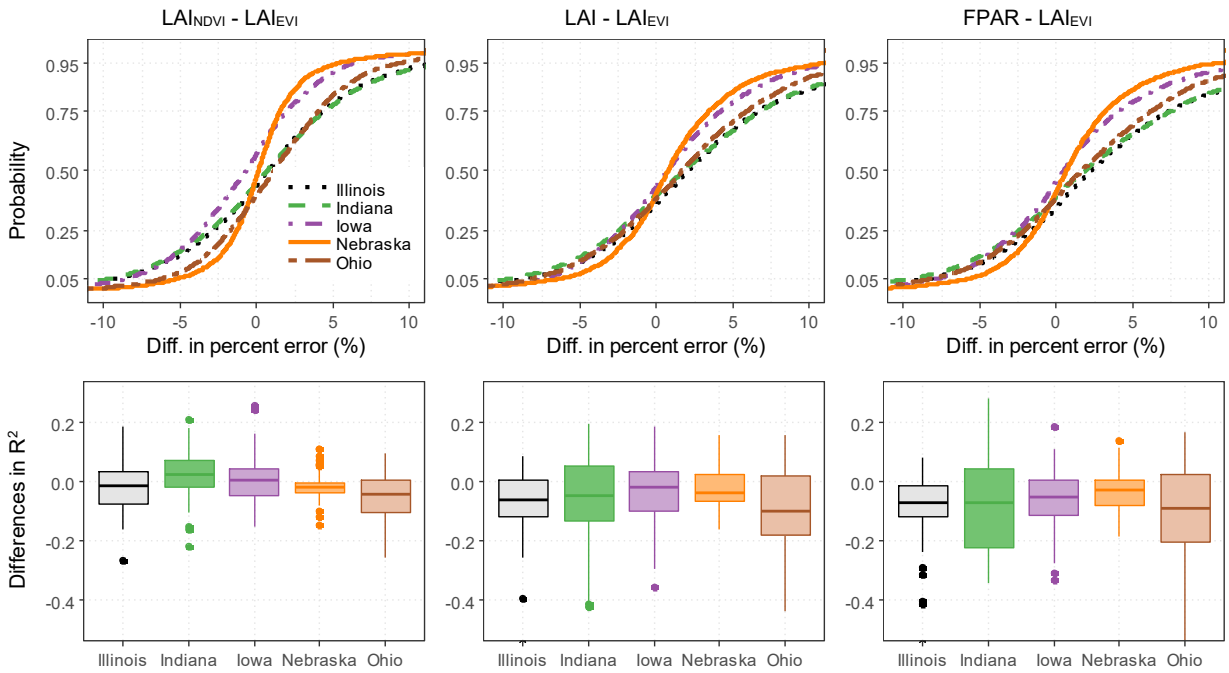

Figure V.9. The top panel shows the cumulative probability distribution of the differences between the percent errors (PE) at the county-level considering the LAI_{NDVI}, LAI and FPAR based forecasts and the PE considering the LAI_{EVI} based forecasts. The bottom panel shows the distribution of the differences in the annual R^2. The comparisons considered the raw composite's temporal resolution and the four spatial domains (except the global domain in Nebraska, which provided exceptionally poor performance). The dots represent outliers in the distribution.

The NDVI-based forecasts provided significantly larger errors in all states except Iowa, where the errors were significantly lower. The results concerning Iowa are particularly interesting as it is precisely the largest producer state in the U.S. While studies have found that the EVI is better than the NDVI for corn yield estimation in the central U.S (Bolton and Friedl, 2013; Johnson et al., 2016), they have made no distinctions among states. As a robustness check, we generated 20 replicas of the annual EVI and NDVI-based forecasts, with each replica considering seven randomly chosen training years. The NDVI based forecasts still provided consistently lower errors in Iowa and higher in the other states. LAI- and FPAR-based forecasts (Fig. V.9) performed worse than EVI-based forecasts, especially in Illinois, Indiana and Ohio. The performance was comparatively less poor in Iowa, which seemed another evidence that this state responds better to the NDVI-based forecasts (as both the MODIS LAI and FPAR are derived from the NDVI).

The differences among products seemed even sharper at the state level (Fig. V.10) compared to the county level. For example, in Indiana and Ohio, the errors using NDVI were on

average about 1.5 times larger than using EVI, while in Iowa they were on average about 20% lower. In Nebraska the NDVI- and EVI-based forecasts provided similar performance.

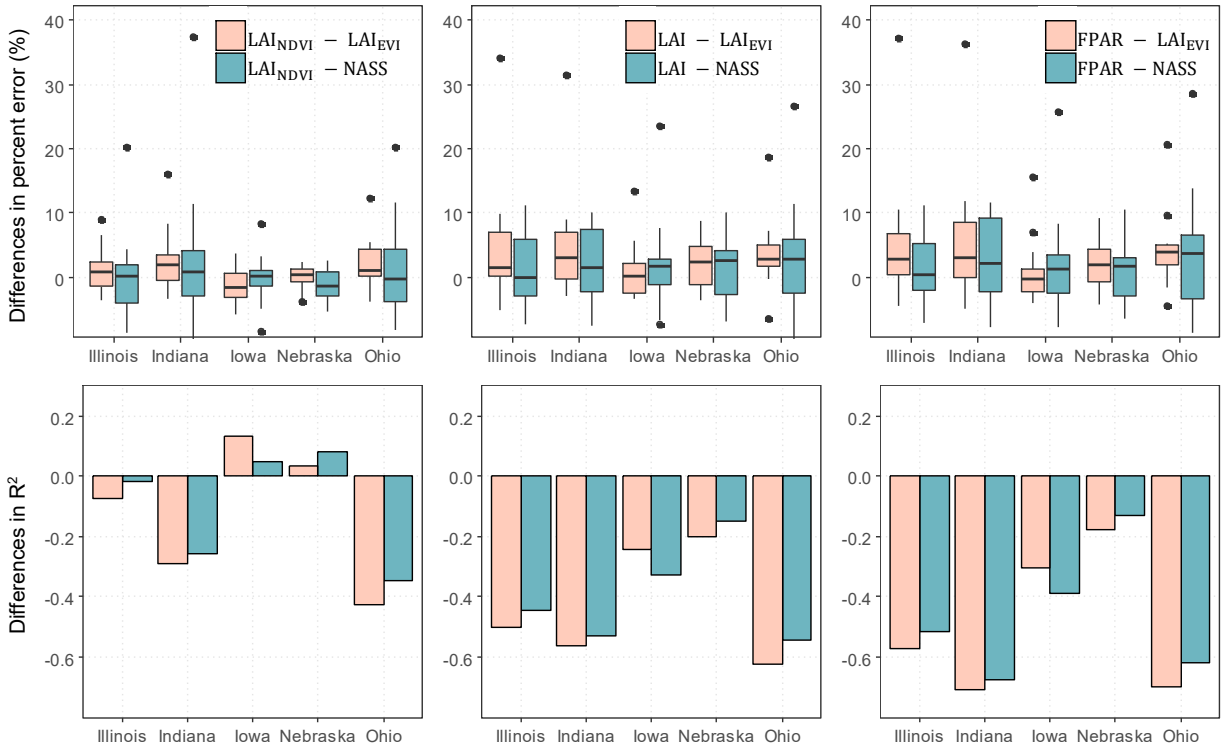

Figure. V.10. Boxplots in the upper panel show the differences in the percent errors (PE) using the LAI_{NDVI}, LAI and FAPAR $\langle\overline{209}\rangle$ with respect to PE using LAI_{EVI} $\langle\overline{209}\rangle$ and the PE using the NASS forecasts. Bar plots in the bottom show the corresponding differences in R^2. The dots represent outliers in the distribution.

Notice that the use of NDVI-based forecasts in Iowa and EVI-based forecasts in the other states provide similar or better performance than the use of the NASS forecasts at any state. A question is whether the model domains that performed best in Iowa considering the EVI-based forecasts perform also best considering the NDVI-based forecast. When developing the same analysis as in Table2 for Iowa, but considering the NDVI-based forecasts, we obtain that the state-based domain performs better than the other domains both at the county and state level, i.e. similarly as when considering the EVI-based forecast.

The performance using the LAI and FPAR based forecasts was remarkably poor in most circumstances compared to the performance using both the EVI based and the NASS forecasts. Based on these results, the choice of any of these biophysical parameters, the LAI or the FPAR, for forecasting purposes seemed not as desirable as the EVI or the NDVI.

Figure V.11 shows the percent error map based on model predictions with the best model schemes. The Figure suggests that errors have notable spatial patterns, with well-defined areas of low and high biases, but tend to be random in time.

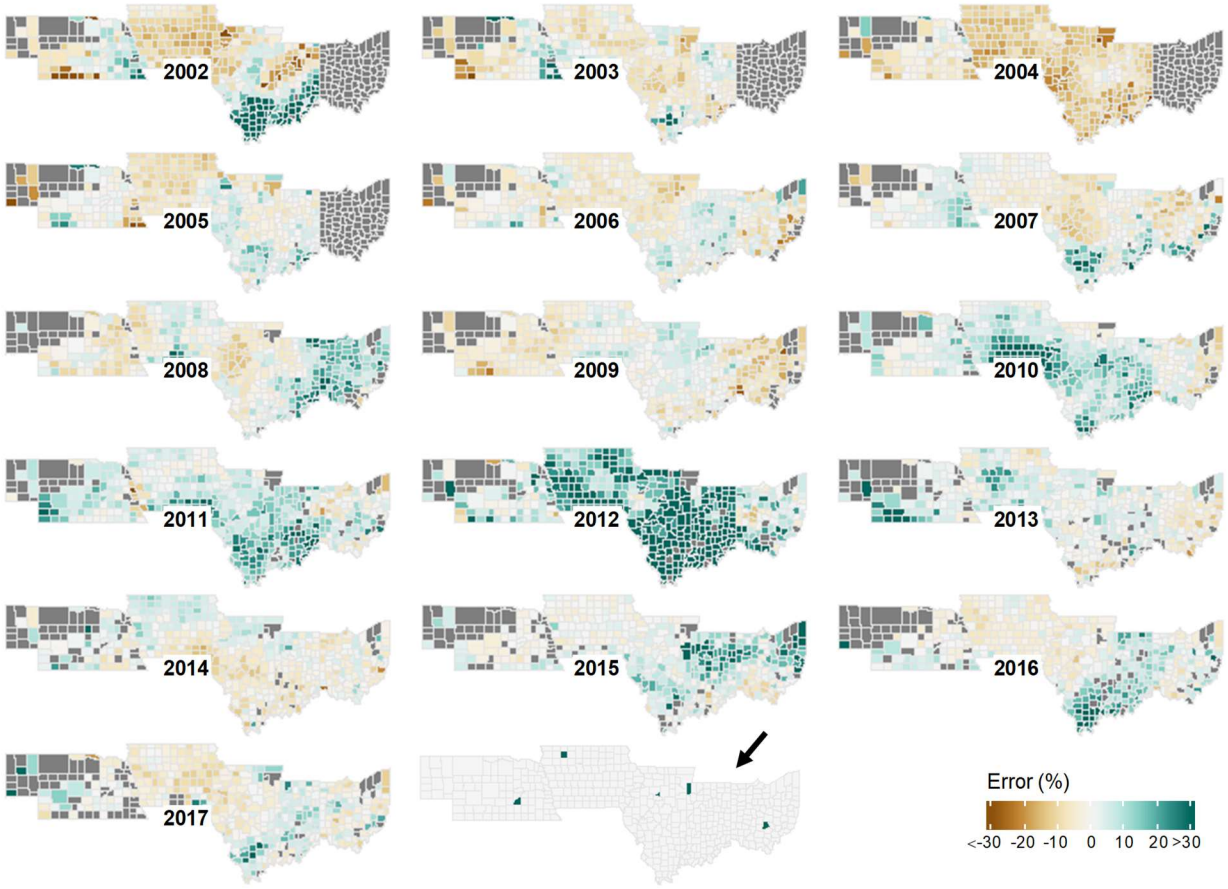

Figure V.11. Error map based on linear regression on $\langle\overline{209}\rangle$ considering the best MODIS products (NDVI for Iowa and EVI for the rest and the best domains at county-level (based on Table V.2: **Global** in Illinois, Indiana, and Iowa, **County** in Nebraska and **District** in Ohio). Polygons in dark gray indicate missing values. The map indicated with an arrow shows the county with the lowest mean percent error across years in each state.

We included a map in Figure V.11 showing the location of the county whose annual mean percent error was the lowest in each state: Putnam in Illinois, Lake in Indiana, Clay in Iowa, Hamilton in Nebraska and Fairfield in Ohio. Their correlations with the true county yields varied between 0.82 in Clay (Iowa) and 0.93 in Fairfield (Ohio). They have a lower correlation (<0.79) with the state yields, except in Putnam (Illinois), where the values coefficients were very similar (~0.83).

4 DISCUSSION

4.1 Best EVI-based forecasting framework

This study found that among the machine learning techniques, the feature strategies and temporal resolutions tested, one of the simplest configurations, involving linear regression on a single original (i.e. untreated for accounting for the actual pixel days) composite of EVI was among the best. This approach captured the site-to-site differences in yields similarly or better than most others did both at the county and state levels. In the following, we discuss probable reasons for this result.

First, regarding the machine learning techniques and the subsetting strategy, a probable cause for the success of the simple approach is that the earlier composite members may capture little or no additional information about the yield variance compared with information available in the last member tested. This aligned with the reports of Labus et al. (2002) who found that early season NDVI from the Advanced Very High Resolution Radiometer (AVHRR) were not good indicators of wheat yields. It may suggest that the production system in the region is resilient to natural stressors, i.e. it can recover from adverse effects potentially occurring at early development stages. The collinearity between time series of MODIS composites may lead to incorrect model identification using elastic net and random forest with multiple composites. Dorman et al., 2014 found that practically no approaches can efficiently deal with the collinearity issues when the correlation between predictors is too high (> 0.7). Another probable reason is that the underlying relationship between the LAI_{EVI} and the yields is truly linear, such that the nonlinear regressions with random forest are outperformed by linear regressions. Our results suggested that the machine learning technique is not the main limiting factor given the available predictors and that new information may be needed for consistently improving the performance with the simplest approach adopted.

Second, regarding the effect of the MODIS composites' timing, the gain by considering the actual pixel days is marginal likely due to the losses in accuracy through the LAI interpolation. Based on this experiment the interpolations with the logistic regression are less accurate, precisely for the period of maximum LAIs, because it tends to over-smooth the variability of time series. Future studies may try to address this by testing other interpolation algorithms (e.g, Jonsson and Eklundh, 2002), or through the combination of spatial and temporal approaches (Borak and

Jasinski, 2009). We found it difficult to establish effective rules (constraints) for guiding the interpolation on that period because of the considerable variability in the growing dynamics.

The model domain significantly affected the performance in Nebraska. In this state, the forecast considering the global domain performed considerably worse than considering the other domains. It means that the model trained over areas with a rainfed regime is inaccurate when tested over the areas under irrigated regimes, most likely because the water stress unevenly affects the foliar development and the yields. In this state, the county-based domain performed consistently better than others at the county level, suggesting that even local differences in the irrigation regime may affect model parameterization. In states other than Nebraska, the response was less sensitive to the chosen domain. Nonetheless, a close inspection the performance considering linear regression showed that some domains tend to be consistently better across states than others are, with the global and county-based domains commonly showing the highest rankings at the county level, and the district- and state-based domains showing the highest rankings at the state level. Notice that based on the results in Nebraska, the performance of the global domain over the other states could be improved by excluding the data from that state from the training.

4.2 Comparison of MODIS products

We found that EVI-based forecast performed considerably better than NDVI-based forecasts in Illinois and especially Indiana and Ohio, while they performed similarly or slightly better in Nebraska and consistently worse in Iowa. These differences in the NDVI-based forecast performance relative to the EVI-based are presumably caused by the variable incidence of errors associated with mixed pixel land uses, to which the NDVI is considerably more sensitive than the EVI (e.g. Gao et al., 2000). Notice that the performance of the NDVI-based forecasts tends to be relatively low over the states where the correlations between the LAI and yields are also low (Fig. V.2), most likely due to the mixed-pixel effects. Similarly, in the previous studies, Sakamoto et al. (2013) showed that the mixed-pixel effects were the main factor affecting the performance of corn yield forecasts based on (a non-linear transformation of) the MODIS NDVI. Seo et al. (2019) monitored crop growth and phenology over Iowa and Illinois with MODIS NDVI and reported large errors over a few counties of Illinois because of mixed land uses.

The LAI and FPAR products at 500 m resolution provided considerably worse performance than the LAIs generated with simple relationships (Eq. 5 and 8) based on the EVI and the NDVI

at 250 m resolution. Based on previous studies (e.g., Chen et al., 2006) the differences in the resolution are unlikely to be the main cause leading to these. Therefore, the operational method used for estimating the LAI and FPAR might be a more probable cause leading to relatively poor performance using these products.

Our results demonstrate that states may respond differently to the different forecasting strategies, mostly because of the variable management systems, environment and land use classification quality across states. The states account for not only different water management conditions but also different cover crop strategies (Singer et al., 2007), planting dates, weather anomalies, pathogen infestations, soils and technologies (e.g., Tannura et al., 2008), which affect the relationship between the canopy quantity and quality and the yields.

4.3 County- versus state-level forecasts

In agreement with previous studies (e.g. Lobell and Burke, 2010) the state-level forecasts aggregated from the county-level mostly provided lower errors than the county level forecasts, suggesting that the discrepancies among counties tend to be unsystematic, so they may partially offset through the aggregation. The R^2 at the state-level tended to be also similar or better if considering the best MODIS products. However, if relying uniquely on the EVI-based forecasts, the state-level R^2 at Iowa was poor, while if relying uniquely upon the NDVI-forecasts, the state-level R^2 at Indiana and Ohio were considerably poor. In general, the improvements through the aggregation were more consistent when using linear regression and elastic net models than random forest models, suggesting that random forest provides not only larger but also more systematic county-level errors. An interesting finding was that the domains that work the best at the county level are not generally the same as the domains that work the best at the state level. For example, while at the county level the global and county domains seemed the best in Indiana and Nebraska, respectively, at the state level the district domain performed best in both states. It seemed that the advantages of the extreme domains (i.e. the size of the training dataset with the global domain, and the ability to better representing the behavior of the individual units with the county domain) are better capitalized at the county level, while the advantages of the intermediate (district- and state-based) model domains are better capitalized at the state level.

4.4 Performance in 2012, the extreme drought year

The performance of the different schemes in extreme years can be of interest, especially in light of the potential impacts of the ongoing climate change on maize yields over the Corn Belt (e.g. Urban et al., 2012). The variability patterns of the LAI and yield in 2012 were completely anomalous (see Fig. V.4), because of the intense drought, and this resulted in large discrepancies between observations and forecasts, mostly by overestimation. Tannura et al. (2008) reported similar issues when testing corn and soybean yield forecasts in another extreme year. While the performance was in general deficient for most schemes, it tended to be considerably poorer when based on regional domains (global and state) and the random forest technique, than when based on other domains or techniques. In Illinois, both the random forest technique and the global domain led to errors that were more than twice as large as the errors from the linear regression technique based on the district- or the county-based domains. Even small inconsistencies in the forecasting model could make the nonlinear approach especially vulnerable to inaccuracies when the range of the testing data is different from the range of the training data, as in 2012. On the other hand, the impact of the agro-environmental variability (e.g. the soil variability) on yield patterns is intuitively stronger in extremes years, which makes the regional domains especially unsuitable. Nonetheless, notice that unlike the errors, the R^2 behaved normally, indicating that most tested schemes discriminated well between areas of smaller and higher yields, while largely overestimating them. Therefore, the range of variability provided by this extreme year should be very helpful in the face of events with similar characteristics in the future, for which the forecasts should perform substantially better.

4.5 Future work

Future research may include sub-seasonal to seasonal forecasts (of, for example, monthly temperatures and precipitation) into the crop yield forecasting frameworks. The weather condition during the reproductive period is a determinant of the final yield. For example, studies (e.g. Thompson, 1969) show that below-average temperatures in August benefit corn productivity over the Corn Belt. Also, there have been considerable improvements in the sub-seasonal to seasonal forecasting from numerical weather predictions during the last years (e.g., Tian et al., 2017). Therefore, improvements in crop yield assessments may arise by combining the MODIS data with

the sub-seasonal to seasonal forecasts within the machine-learning frameworks. The sensitivity of results to explicitly accounting for spatio-temporal dependencies between data points also deserves more research (e.g., You et al., 2017). The spatial variability of soil properties, weather and crop management practices can affect the spatial yield patterns in complex ways. The error patterns observed in the study showed a strong spatial structure, but they tended to be unsystematic over time, which may make it difficult to account for the variability. Comprehensive studies are also needed to evaluate the potential of recent approaches based on deep learning architectures (e.g., Kuwata and Shibasaki 2015; You et al., 2017). Based on the results of You et al., (2017), the gains of considering these modern tools instead of more traditional techniques may be considerably larger for the forecasts at the late stages of the crop season than at early and mid-stages, and some deep learning architectures can perform even worse.

5 CONCLUSIONS

The accurate forecasting of corn productivity over the Corn Belt has large implications worldwide. While previous studies have demonstrated the usefulness of MODIS composites for this purpose, few attempts have examined the response of the corn yield forecasts to the different MODIS products, model domains, machine-learning techniques, product subsets, and temporal resolution of the pixel composites. Little is also known about if the schemes providing the best performance at the county level are also optimal when aggregating the forecasts to the state level. This study for the first time developed an optimized MODIS-based framework for mid-season corn yield forecasting at the county and state levels over five major producers states that considered multiple MODIS products, model domains, machine learning techniques, product subsets, and temporal resolutions. The optimized forecasting framework outperformed the NASS forecasts issued in August. We considered time series of operational 16-day composites of EVI and NDVI at 250 m resolution and 8-day composites of LAI and FPAR at 500 m resolution issued between early (end of May) and mid-season (end of July) as our primary MODIS products used for yield forecasting.

We find that the forecasts based on linear regressions with the latest mid-season EVI composite mostly showed better performance than the random forest- or elastic-net based forecasts. In particular, the forecasts with the random forest technique were consistently worse, especially when evaluated at the state level. The forecast schemes involving EVI composites from

the early season usually perform worse than the schemes solely considering information from mid-season. While the choice of the model domain has a significant impact in Nebraska, where the agricultural lands are both rainfed and irrigated, it has a more limited impact in the other states. The global and county-based domains often showed better forecasts at county-level, while the district- and state-based domains provided better state-level forecasts. The models based on global and state-based domains led to considerably poor forecasts in 2012, and therefore caution should be taken when tested on years affected by extreme events. The forecasts respond similarly to the temporal resolutions tested, including a daily (high) resolution considering the true day of every pixel composite and the original coarse resolution of the MODIS products, which indicates that the loss in accuracy induced by treating the datasets for accounting for the true day offset the potential gain of increasing the resolution. When using simple linear regression based on the latest composite product, the EVI based forecasts perform the best in Illinois, Indiana and Ohio, while the NDVI based forecasts perform the best in Iowa, while they perform similarly in Nebraska. The study shows that both the error and the R^2 at the state level are usually better than the errors of the state level, provided we use the best MODIS products across states. The LAI and FPAR based forecasts are instead worse than the VI-based forecasting in most circumstances and, in occasions, completely inaccurate. An important finding in this study is that the forecast performance and the response to changes in model configuration are variable among states, suggesting that the forecasting strategies should be state-oriented. In summary, our results indicate that simple adjustments in the forecasting framework can have a large impact on forecasting performance.

REFERENCES

1. Becker-Reshef, II., Vermote, E., Lindeman, M., Justice, C., 2010. A generalized regression-based model for forecasting winter wheat yields in Kansas and Ukraine using MODIS data. Remote sensing of environment. 114(6), pp.1312-1323.
2. Bolton, D.K., Friedl, M.A., 2013. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. Agricultural and Forest Meteorology. 173, pp.74-84.
3. Borak, J.S., Jasinski, M.F., 2009. Effective interpolation of incomplete satellite-derived leaf-area index time series for the continental United States. Agricultural and Forest Meteorology. 149(2), pp.320-332.
4. Breiman, L., 2001. Random forests. Machine Learning. 45(1): 5–32, 2001. 18

5. Chen, P.Y., Fedosejevs, G., Tiscareno-Lopez, M., Arnold, J.G., 2006. Assessment of MODIS-EVI, MODIS-NDVI and VEGETATION-NDVI composite data using agricultural measurements: An example at corn fields in western Mexico. Environmental monitoring and assessment. 119(1-3), pp.69-82.
6. Delucchi, L., Neteler, M., 2013. pyModis: the Python library for MODIS data. In: FOSS4G 2013, Nottingham, 17-21 September 2013. url: http://conf.lucadelu.org/pymodis/ handle: http://hdl.handle.net/10449/22573.
7. Didan, K., 2015. MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V006 [Data set]. NASA EOSDIS LP DAAC. doi: 10.5067/MODIS/MOD13Q1.006.
8. Doraiswamy, P.C., Sinclair, T.R.; Hollinger, S., Akhmedov, B., Stern, A., Prueger, J., 2005. Application of MODIS derived parameters for regional crop yield assessment. Remote Sens. Environ. 97, 192–202.
9. Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J., Münkemüller, T., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography. 36(1), pp.27-46.
10. FAO, 2019. FAOSTAT statistical database. Available at: http://www.fao.org/faostat/en/
11. Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. 33(1), 1-22.
12. Funk, C., Budde, M.E., 2009. Phenologically-tuned MODIS NDVI-based production anomaly estimates for Zimbabwe. Remote Sensing of Environment. 113(1), pp.115-125.
13. Gao, X., Huete, A.R., Ni, W., Miura, T., 2000. Optical–biophysical relationships of vegetation spectra without background contamination. Remote Sensing of Environment. 74(3), pp.609-620.
14. GDAL/OGR contributors, 2018. GDAL/OGR geospatial data abstraction software library. Open Source Geospatial Foundation.
15. Guindin-Garcia, N., Gitelson, A.A., Arkebauer, T.J., Shanahan, J., Weiss, A., 2012. An evaluation of MODIS 8-and 16-day composite products for monitoring maize green leaf area index. Agricultural and Forest Meteorology. 161, pp.15-25.
16. Hansen, J.W., Jones, J.W., 2000. Scaling-up crop models for climate variability applications. Agricultural Systems. 65(1), pp.43-72.

17. Hochheim, K.P., Barber, D.G., 1998. Spring wheat yield estimation for Western Canada using NOAA NDVI data. Canadian journal of remote sensing. 24(1), pp.17-27.
18. Horie, T., Yajima, M., Nakagawa, H., 1992. Yield forecasting. Agricultural systems. 40(1-3), pp.211-236.
19. Jaafar, H.H., Ahmad, F.A., 2015. Crop yield prediction from remotely sensed vegetation indices and primary productivity in arid and semi-arid lands. International Journal of Remote Sensing, 36(18), pp.4570-4589.
20. Johnson, D.M., 2014. An assessment of pre-and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. Remote Sensing of Environment. 141, pp.116-128.
21. Johnson, D.M., 2016. A comprehensive assessment of the correlations between field crop yields and commonly used MODIS products. International Journal of Applied Earth Observation and Geoinformation. 52, pp.65-81.
22. Johnson, M.D., Hsieh, W.W., Cannon, A.J., Davidson, A., Bédard, F., 2016. Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods. Agricultural and forest meteorology. 218, pp.74-84.
23. Johnson, R.A., Wichern, D.W., 2002. Applied multivariate statistical analysis (Vol. 5, No. 8). Upper Saddle River, NJ: Prentice Hall.
24. Jonsson, P., Eklundh, L., 2002. Seasonality extraction by function fitting to time-series of satellite sensor data. IEEE transactions on Geoscience and Remote Sensing. 40(8), pp.1824-1832.
25. Kogan, F., Kussul, N., Adamenko, T., Skakun, S., Kravchenko, O., Kryvobok, O., Shelestov, A., Kolotii, A., Kussul, O., Lavrenyuk, A., 2013. Winter wheat yield forecasting in Ukraine based on Earth observation, meteorological data and biophysical models. International Journal of Applied Earth Observation and Geoinformation. 23, pp.192-203.
26. Kuwata, K., Shibasaki, R. 2015. Estimating crop yields with deep learning and remotely sensed data. In 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). 858–861. IEEE.
27. Labus, M.P., Nielsen, G.A., Lawrence, R.L., Engel, R., Long, D.S., 2002. Wheat yield estimates using multi-temporal NDVI satellite imagery. International Journal of Remote Sensing. 23(20), pp.4169-4180.

28. Liaw, A., M. Wiener, 2002. Classification and Regression by randomForest. R News 2(3), 18-22.
29. Lobell, D.B., Burke, M.B., 2010. On the use of statistical models to predict crop yield responses to climate change. Agricultural and forest meteorology. 150(11), pp.1443-1452.
30. Lobell, D.B., Roberts, M.J., Schlenker, W., Braun, N., Little, B.B., Rejesus, R.M., Hammer, G.L., 2014. Greater sensitivity to drought accompanies maize yield increase in the US Midwest. Science. 344(6183), pp.516-519.
31. Mallya, G., Zhao, L., Song, X.C., Niyogi, D., Govindaraju, R.S., 2013. 2012 Midwest drought in the United States. Journal of Hydrologic Engineering. 18(7), pp.737-745.
32. Mkhabela, M.S., Bullock, P., Raj, S., Wang, S., Yang, Y., 2011. Crop yield forecasting on the Canadian Prairies using MODIS NDVI data. Agricultural and Forest Meteorology. 151(3), pp.385-393.
33. Mkhabela, M.S., Mkhabela, M.S., Mashinini, N.N., 2005. Early maize yield forecasting in the four agro-ecological regions of Swaziland using NDVI data derived from NOAA's-AVHRR. Agricultural and Forest Meteorology. 129(1-2), pp.1-9.
34. Montgomery, D.C., Peck, E.A., Vining, G.G., 2012. Introduction to linear regression analysis (Vol. 821). John Wiley & Sons.
35. Myneni, R., Knyazikhin, Y., Park, T., 2015. MOD15A2H MODIS/Terra Leaf Area Index/FPAR 8-Day L4 Global 500m SIN Grid V006 [Data set]. NASA EOSDIS Land Processes DAAC. doi: 10.5067/MODIS/MOD15A2H.006
36. R Core Team, 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
37. Rembold, F., Atzberger, C., Savin, II., Rojas, O., 2013. Using low resolution satellite imagery for yield prediction and yield anomaly detection. Remote Sensing. 5(4), pp.1704-1733.
38. Sakamoto, T., Gitelson, A.A., Arkebauer, T.J., 2013. MODIS-based corn grain yield estimation model incorporating crop phenology information. Remote Sensing of Environment. 131, pp.215-231.
39. Seo, B., Lee, J., Lee, K.D., Hong, S., Kang, S., 2019. Improving remotely-sensed crop monitoring by NDVI-based crop phenology estimators for corn and soybeans in Iowa and Illinois, USA. Field Crops Research. 238, pp.113-128.

40. Son, N.T., Chen, C.F., Chen, C.R., Chang, L.Y., Duc, H.N., Nguyen, L.D., 2013. Prediction of rice crop yield using MODIS EVI− LAI data in the Mekong Delta, Vietnam. International Journal of Remote Sensing, 34(20), pp.7275-7292.
41. Singer, J.W., Nusser, S.M., Alf, C.J., 2007. Are cover crops being used in the US corn belt?. Journal of Soil and Water Conservation. 62(5), pp.353-358.
42. Tannura, M.A., Irwin, S.H., Good, D.L., 2008. Weather, technology, and corn and soybean yields in the US corn belt. Technology, and Corn and Soybean Yields in the US Corn Belt (February 1, 2008).
43. Thompson, L.M., 1969. Weather and Technology in the Production of Corn in the US Corn Belt 1. Agronomy Journal. 61(3), pp.453-456.
44. Tian, D., Wood, E.F., Yuan, X., 2017. CFSv2-based sub-seasonal precipitation and temperature forecast skill over the contiguous United States. Hydrology and Earth System Sciences. 21(3), pp.1477-1490.
45. Urban, D., Roberts, M.J., Schlenker, W., Lobell, D.B., 2012. Projected temperature changes indicate significant increase in interannual variability of US maize yields. Climatic change. 112(2), pp.525-533.
46. Vogel, F.A., Bange, G.A., 1999. Understanding Crop Statistics. Frederic A. Vogel, U.S. Department of Agriculture. Miscellaneous Publication No. 1554.
47. Wardlow, B.D., Egbert, S.L., 2008. Large-area crop mapping using time-series MODIS 250 m NDVI data: An assessment for the US Central Great Plains. Remote sensing of environment. 112(3), pp.1096-1116.
48. You, J., Li, X., Low, M., Lobell, D., Ermon, S., 2017. Deep gaussian process for crop yield prediction based on remote sensing data. In Thirty-First AAAI Conference on Artificial Intelligence.
49. Zhang, X., Friedl, M.A., Schaaf, C.B., Strahler, A.H., Hodges, J.C., Gao, F., Reed, B.C., Huete, A., 2003. Monitoring vegetation phenology using MODIS. Remote sensing of environment. 84(3), pp.471-475.
50. Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology). 67(2), pp.301-320.

SUPPLEMENTAL MATERIALS

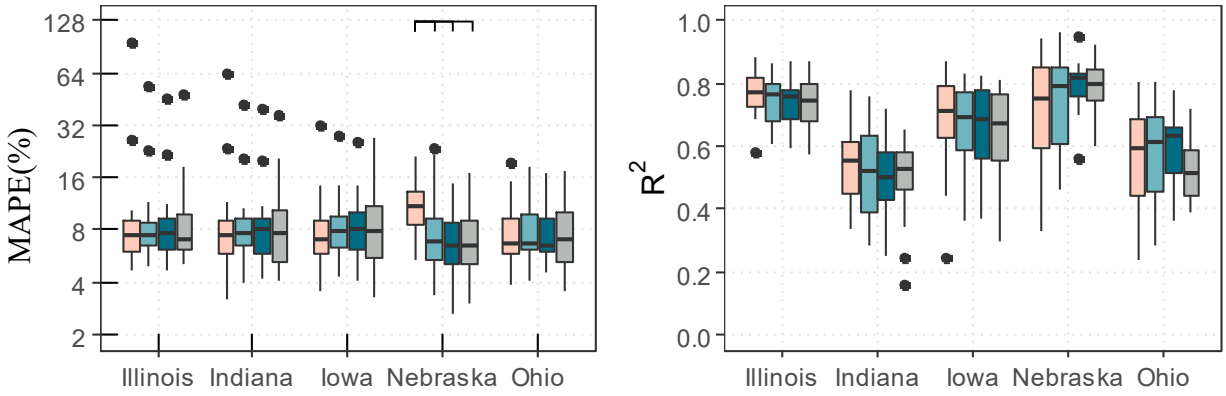

Figure V.A1. Distribution of annual mean absolute percentage error (MAPE) and annual R^2 of the county-level forecasts for the different spatial domains using the linear regression based on ⟨209⟩. The bars in the top of the plots denote significant differences with respect to the domain aligned with the mark on the left. The dots represent outliers in the distribution.

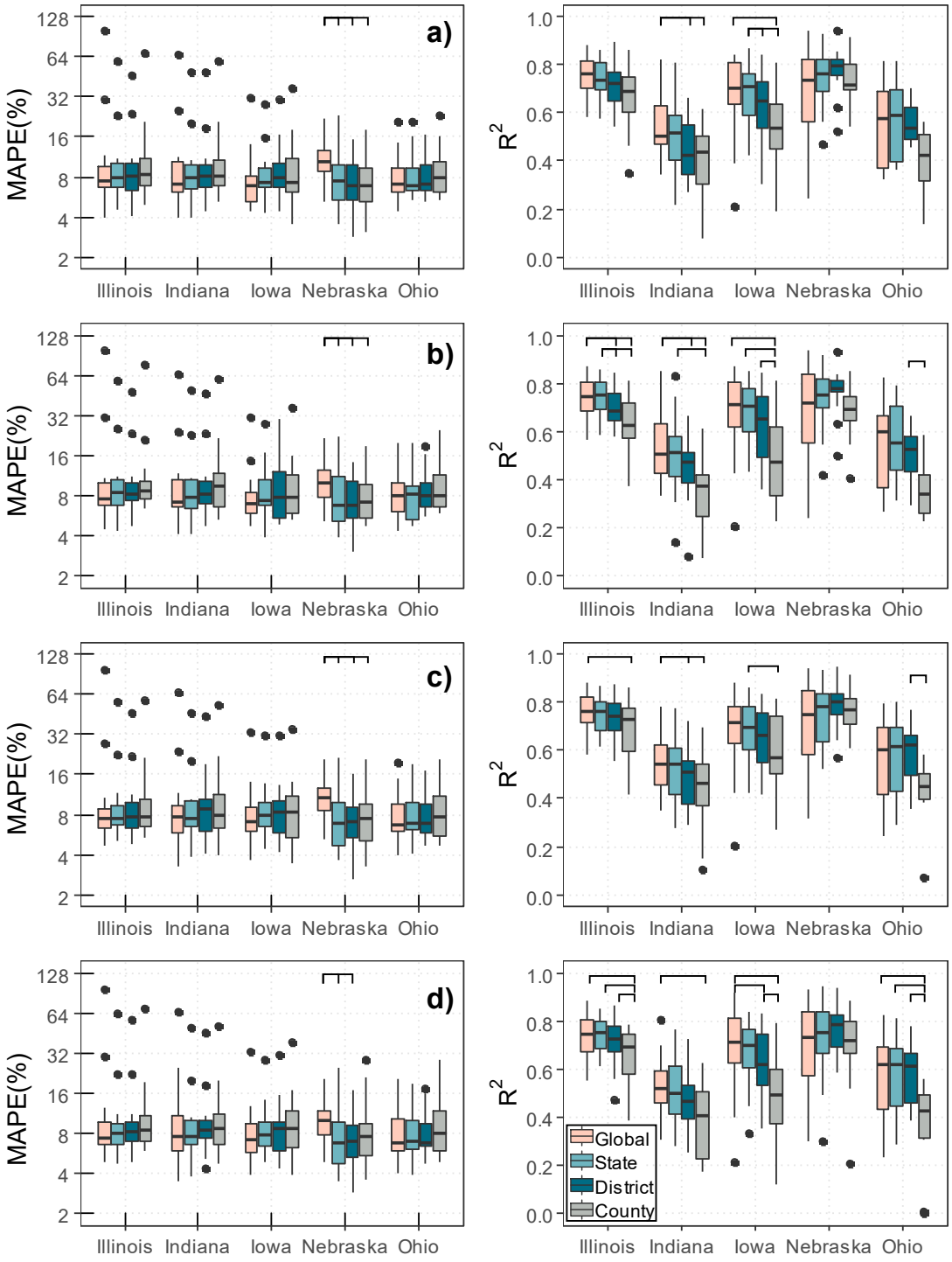

Figure V.A2. Distribution of annual mean absolute percentage error (MAPE) and annual R^2 of the county-level forecasts for the different spatial domains using the elastic net based on a) $\langle \overline{161}, \overline{177}, \overline{193}, \overline{209} \rangle$, b) $\langle \overline{153} \text{ to } \overline{209}, \text{ in steps of } 8 \rangle$, c) $\langle 193, 201, 201 \rangle$ and d) $\langle 153 \text{ to } 209, \text{ in steps of } 8 \rangle$. The bars in the top of the plots denote significant differences with respect to the domain aligned with the mark on the left. The dots represent outliers in the distribution.

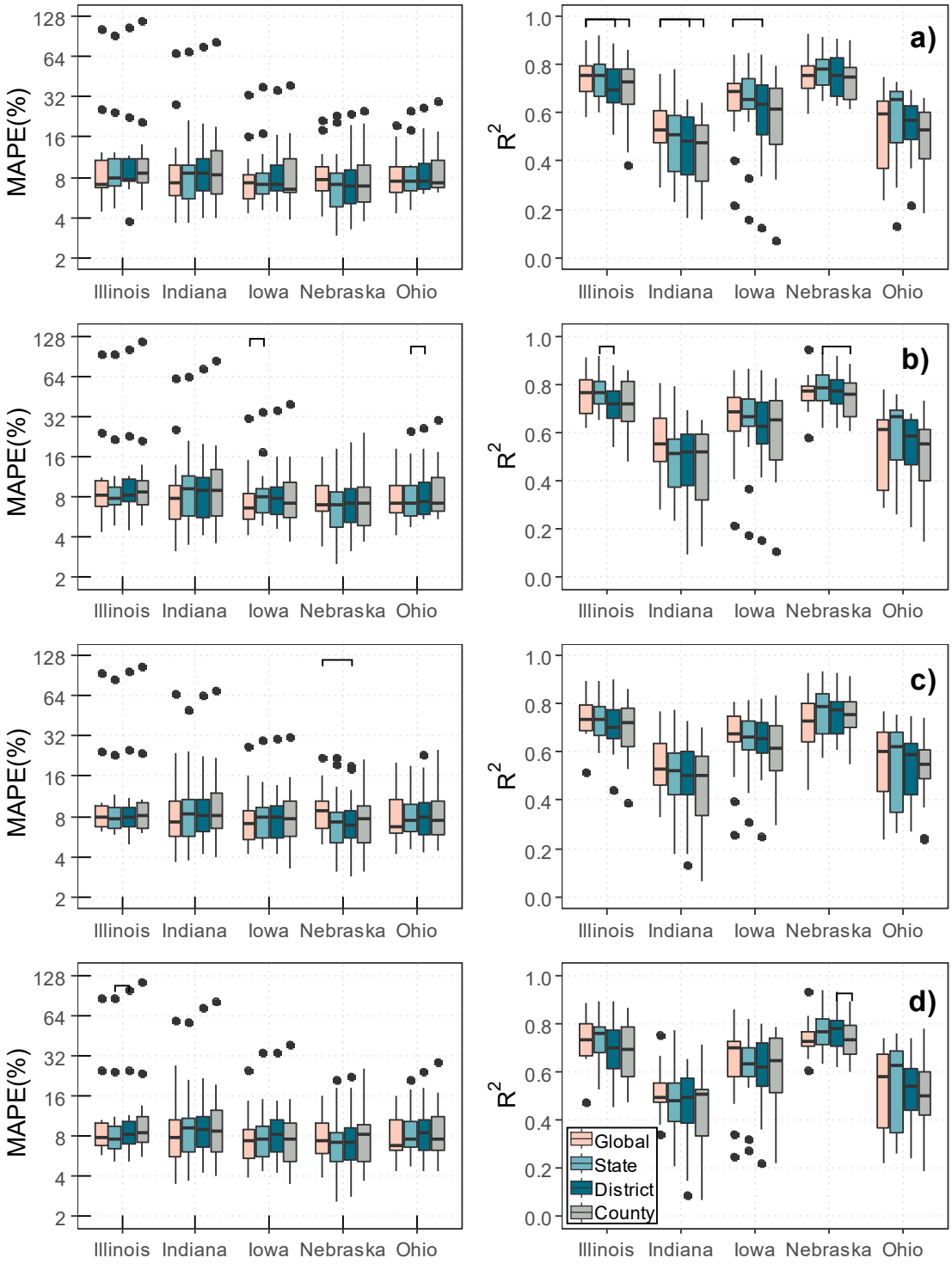

Figure V.A3. Distribution of annual mean absolute percentage error (MAPE) and annual R^2 of the county-level forecasts for the different spatial domains using the random forest based on a) $\langle \overline{161}, \overline{177}, \overline{193}, \overline{209} \rangle$, b) $\langle \overline{153} \text{ to } \overline{209}, \text{ in steps of } 8 \rangle$, c) $\langle 193, 201, 201 \rangle$ and d) $\langle 153 \text{ to } 209, \text{ in steps of } 8 \rangle$. The bars in the top of the plots denote significant differences with respect to the domain aligned with the mark on the left. The dots represent outliers in the distribution. The dots represent outliers in the distribution.

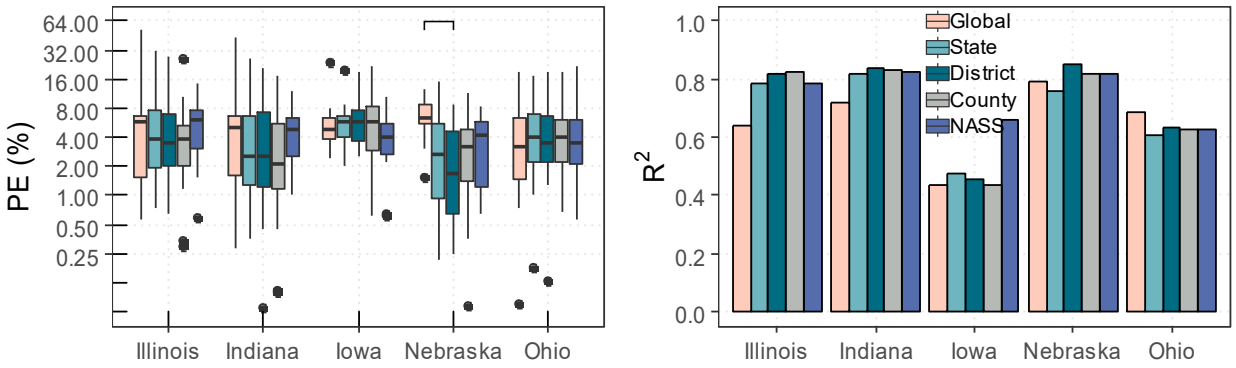

Figure V.B1. Distribution of percent error (PE) and R^2 of the annual **state-level** forecasts for the different spatial domains using the linear regression based on $LAI_{EVI}\langle 209 \rangle$ and the PE and R^2 with the NASS forecasts. The bars in the top of the plots denote significant differences with respect to the domain aligned with the mark on the left. The dots represent outliers in the distribution.

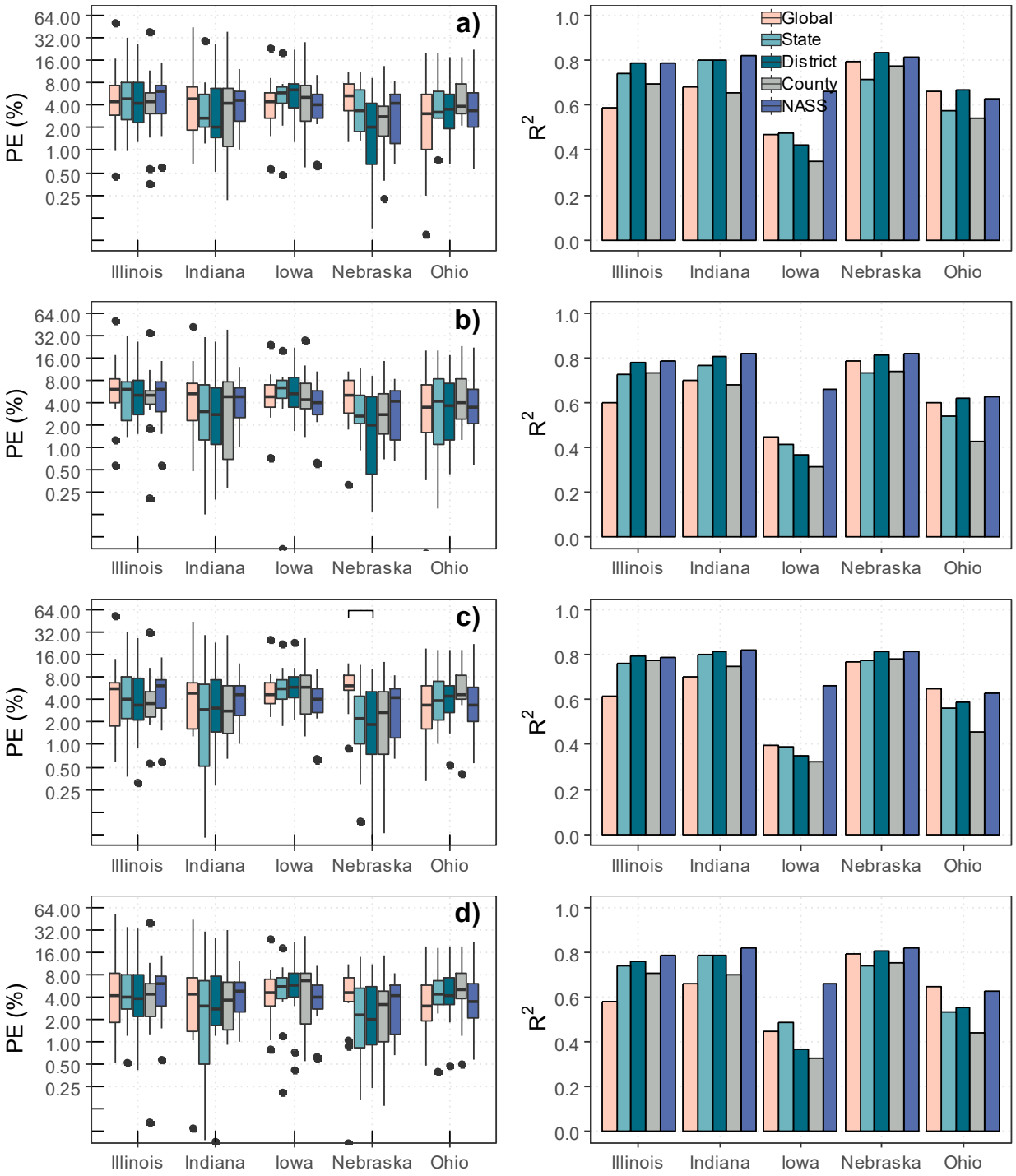

Figure V.B2. Distribution of percent error (PE) and R^2 of the annual **state-level** forecasts for the different spatial domains using the elastic net with the LAI_{EVI} based on a) $\langle \overline{161}, \overline{177}, \overline{193}, \overline{209} \rangle$, b) $\langle \overline{153} \text{ to } \overline{209}, \text{ in steps of } 8 \rangle$, c) $\langle 193, 201, 209 \rangle$ and d) $\langle 153 \text{ to } 209, \text{ in steps of } 8 \rangle$, as well as the PE and R^2 with the NASS forecasts. The bars in the top of the plots denote significant differences with respect to the domain aligned with the mark on the left. The dots represent outliers in the distribution.

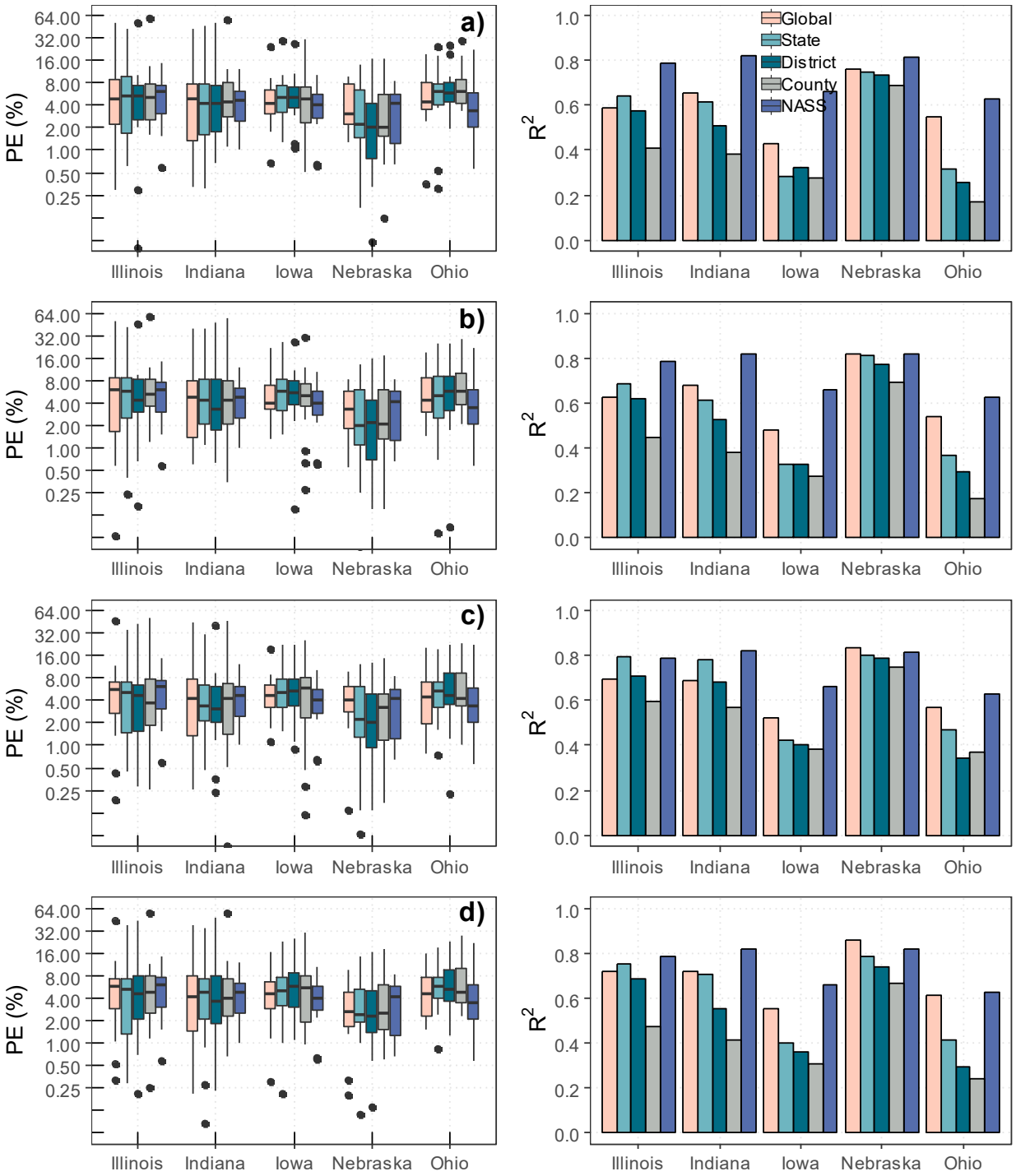

Figure V.B3. Distribution of percent error (PE) and R^2 of the annual **state-level** forecasts for the different spatial domains using the random forest with the LAI_{EVI} based on a) $\langle \overline{161}, \overline{177}, \overline{193}, \overline{209} \rangle$, b) $\langle \overline{153}$ to $\overline{209}$, in steps of $8 \rangle$, c) $\langle 193, 201, 209 \rangle$ and d) $\langle 153$ to 209, in steps of $8 \rangle$ as well as the PE and R^2 with the NASS forecasts. The dots represent outliers in the distribution.

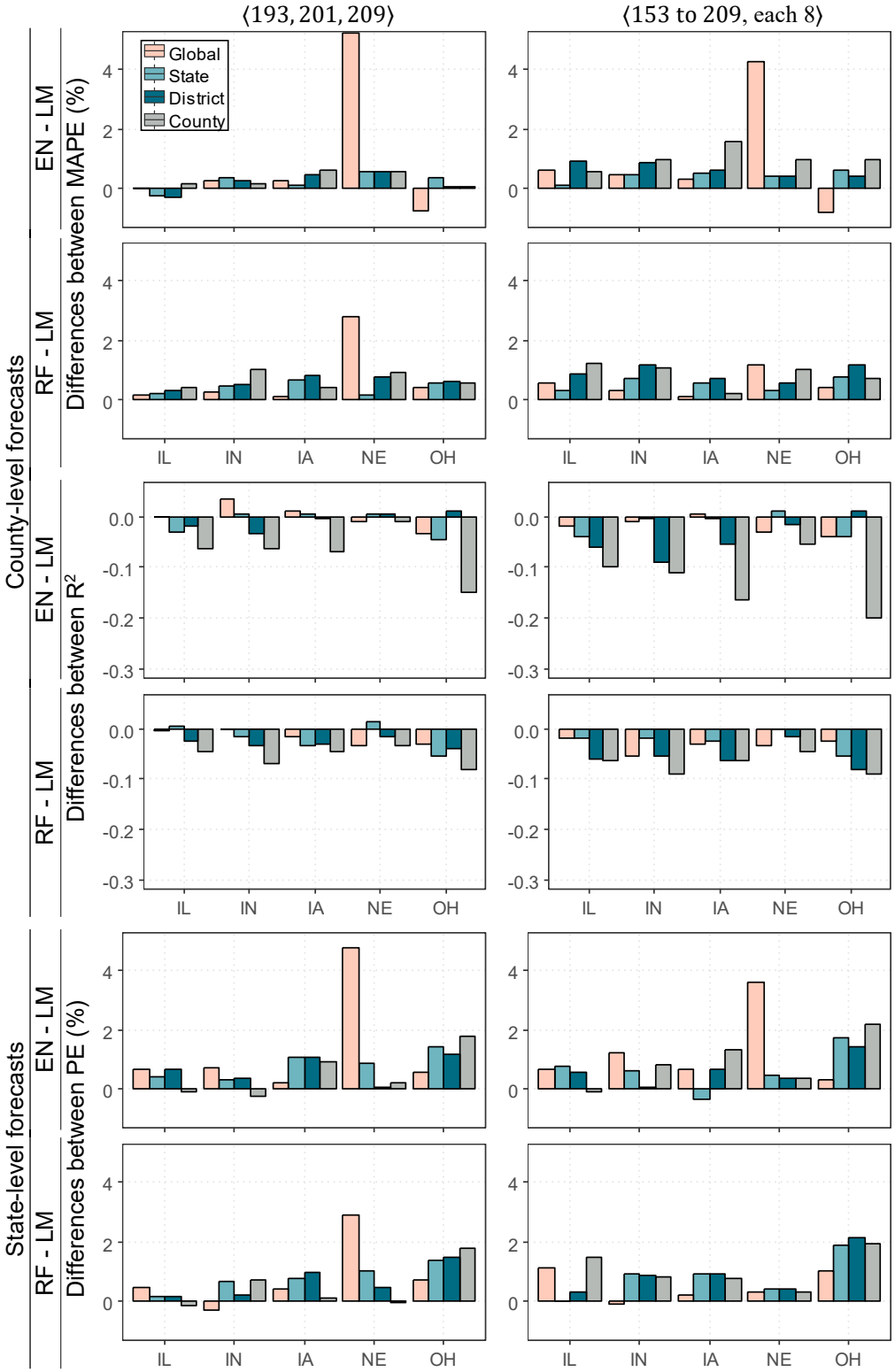

Figure V.C1. Idem to Fig. V.7, but for elastic net and random regressions based on subsets of the LAI_{EVI} (indicated on the top of the figure) at high temporal resolution.

CHAPTER VI: CONCLUDING REMARKS AND FUTURE WORK

This study evaluated novel data-driven approaches for improving water management and crop yield forecasting using NWP forecasts, remote sensing data and machine learning techniques.

Specifically, the study has evaluated the ability of leading NWP models for the medium-range ensemble forecasting of daily and weekly ET_0 (daily rainfall) over the CONUS (Brazil), finding that the ECMWF ensemble forecasts are considerably better than NCEP forecasts, and (in the case of ET_0) also better than the UKMO forecasts. For the first time, multi-model ensemble forecast of daily and weekly ET_0 were evaluated and compared against single model ensemble forecasts, showing that the multi-model combining the ECMWF and the UKMO forecasts leads to improvements at short lead times (1-2 days) and over the southern and western regions of the U.S. compared with the best single model forecasts (using ECMWF). This research represents a step forward in the use of probabilistic techniques for the post-processing of precipitation and ET_0 ensemble forecasts, by evaluating several AF methods and a LR method for the rainfall post-processing, as well as the NGR, the BMA and the AFK methods for ET_0 post-processing. The post-processing in most cases greatly improved the forecast performance, especially when considering the NCEP forecasts and over regions where the performance is comparatively poor, such as near the coast or in areas with complex orography (in the case of the ET_0 forecasts in the U.S.) and over the Amazonia biome (in the case of the rainfall in Brazil). While the post-processing itself of the ensemble forecasts of rainfall was critical for having skillful predictions, the type of the post-processing technique, either the AF or the LR techniques, had little impact on the forecast performance. The NGR, AKD, and BMA post-processing methods were shown to improve the probabilistic performance of both the daily and weekly ET_0 ensemble forecasts compared with a linear regression bias correction method, with the NGR post-processing of the ECMWF and ECMWF–UKMO forecasts providing the most cost-effective ET_0 forecasting.

This study has also shown that several crop yield forecasting approaches with MODIS datasets may be susceptible to improvement just by adopting simple changes in the prediction framework. A surprising finding is that simple linear regression forecasts of the county- and state-level corn yields based on the single latest EVI composite in mid-season are similar or better than the elastic net and random forest forecasts based on multi-temporal composites of EVI. It has been also found that among the MODIS products tested, the EVI provides best forecasts in Illinois,

Indiana and Ohio, but the NDVI provides best forecasts in Iowa, and both the EVI and NDVI provide similar forecasts Nebraska. The LAI and FPAR, instead, commonly provide poor forecasts compared with the EVI or the NDVI-based forecasts. The study demonstrates that the training domains that are best for the county level forecasting, are not the same as the training domains that are best for the state level forecasting. One important finding is that, despite the simplicity of the frameworks identified as the best, they provide forecasts that commonly outperform the National Agricultural Statistical Service (NASS) forecasts.

In summary, the results reliably supported a main hypothesis of this work, by showing that although there is growing availability of information from numerical weather prediction and remote sensing, much research is needed for translating those datasets into products that may support the decision-making in agriculture.

While the study focused on the medium range NWP ensemble forecasts, further research is needed for using sub-seasonal to seasonal forecast information for addressing the needs of water management in agriculture. I envisage that the combination of the MODIS datasets with the sub-seasonal to seasonal forecasts may also help to further improve the crop yield forecasting across sclaes. Efforts are also needed for evaluating the potential of new remote sensing datasets such as the Harmonized Landsat Sentinels (HLS), and new machine learning techniques, including the deep learning architectures, for improving the water management and crop yield forecasting.