Evaluation and Disaggregation of Climate Model Outputs for European Drought Prediction

Dissertation
kumulativ
zur Erlangung des akademischen Grades doctor rerum naturalium
(Dr. rer. nat.)

vorgelegt dem Rat der Chemisch-Geowissenschaftlichen Fakultät der
Friedrich-Schiller-Universität Jena

von Dipl.-Math. Stephan Thober
geboren am 28. August 1984 in Salzwedel

Betreuerin: Prof. Dr. Sabine Attinger
Ko-Betreuer: Dr. Luis E. Samaniego

Diese Dissertation ist während meiner Beschäftigung als wissenschaftlicher Mitarbeiter in der Abteilung für Hydrosystemmodellierung innerhalb des Helmholtz-Zentrums für Umweltforschung - UFZ im Rahmen der WESS (www.wess.info) und REKLIM (www.reklim.de) Projekte entstanden. Diese Dissertation wurde durch die Helmholtz Interdisciplinary Graduate School for Environmental Research (HIGRADE) unterstützt.

Gutachter:

Prof. Dr. Sabine Attinger	Friedrich-Schiller-Universität Jena
Prof. Dr. Karsten Schulze	Universität für Bodenkultur Wien
Prof. Dr. Ralf März	Martin-Luther-Universität Halle-Wittenberg

Die Dissertationsverteidigung fand am 27. Januar 2016 statt.

Für meine Oma

Contents

List of Figures 9

List of Tables 15

Danksagung 17

Übersicht 19

Gesamtzusammenfassung 21

Abstract 27

1 Introduction 33

1.1 Background 35
1.2 Seasonal Drought Prediction System 36
    1.2.1 Evaluation of Multi-Model Ensembles of Climate Models 39
    1.2.2 Downscaling Techniques for Meteorological Variables 41
1.3 Objectives and Research Questions 43

2 Robust Ensemble Selection by Multivariate Evaluation of Extreme Precipitation and Temperature Characteristics 47

2.1 Abstract 49
2.2 Introduction 49
2.3 Data Sets and Study Domain 52
2.4 Methods 54
    2.4.1 Selection of Extreme-Meteorological Indices 54
    2.4.2 The Test Statistic and Hypothesis Testing 55
    2.4.3 Overall Ranking 56
    2.4.4 Rejection Rate 56
    2.4.5 Robust Ensemble Selection 56
2.5 Results and Discussion 59
    2.5.1 Dimensionality of the Input Data 59
    2.5.2 Statistics and Tests 60
    2.5.3 Reliability Ensemble Averaging 67

2.5.4 Model Ranking and Ensemble Performance . . . 68
2.5.5 Implications for Policy Recommendations and Impact Assessments . . . 72
2.6 Summary and Conclusions . . . 74

3 Stochastic Temporal Disaggregation of Monthly Precipitation for Regional Gridded Data Sets 77

3.1 Abstract . . . 79
3.2 Introduction . . . 79
3.3 Method for Seamless Stochastic Temporal Disaggregation . . . 82
3.3.1 Preprocessing and Estimation of $\mathcal{F}$ and $\mathcal{B}$ . . . 84
3.3.2 Generation of Precipitation Realizations . . . 86
3.4 Data Sets and Study Domain . . . 87
3.5 Results and Discussion . . . 87
3.5.1 Estimation of Intensity Classes C and Distributions $\mathcal{F}$ . . . 87
3.5.2 Verification of Precipitation Distribution Functions . . . 91
3.5.3 Probability of Dry Days . . . 95
3.5.4 Extreme Precipitation Characteristics . . . 96
3.5.5 Spatial Correlation for different Scales and Seasons . . . 99
3.5.6 Evaluation at Multiple Temporal Scales . . . 101
3.5.7 Temporal Correlation . . . 103
3.6 Conclusions . . . 104

4 On the Capability of the North American Multi-Model Ensemble for Seasonal Soil Moisture Drought Prediction over Europe 107

4.1 Abstract . . . 109
4.2 Introduction . . . 109
4.3 Methods and Datasets . . . 111
4.3.1 Climate Forecasts (CFs) . . . 111
4.3.2 Construction of Soil Moisture Forecasts . . . 112
4.3.3 Calculation of Soil Moisture Index . . . 113
4.3.4 Subensemble Selection . . . 114
4.4 Results and Discussion . . . 114
4.4.1 Representation of Spatio-Temporal SMI Dynamics and the Effect of Model Averaging . . . 114
4.4.2 Subensemble and Single Model performance for SMI and Drought Forecasts 118
4.4.3 Spatio-Temporal Distribution of Drought Forecasting Skill . . . 119
4.4.4 Forecasting Skill during Drought Development and Recovery . . . 123
4.5 Summary and Conclusions . . . 125

5 Discussion and Conclusions 127

5.1 Summary . . . 129
5.2 Discussion of Spatial Downscaling . . . 130
5.3 Discussion of Subensemble Selection . . . 133
5.4 Discussion of Extreme Indices . . . 134

5.5 Conclusions and Perspectives 136

Appendices **141**

A Anchor sampling **141**

B Equations for ARB and MARB **145**

C Technical Material for "Stochastic Temporal Disaggregation of Monthly Precipitation for Regional Gridded Data Sets" **147**

C.1 Algorithm for the Estimation of the Intensity Classes 147

C.2 Parameter Estimation of Spatial Sampling 148

C.3 Spatial Clustering of CDFs 150

D Equitable Threat Score **153**

E Drought severity and area **155**

Bibliography **157**

Selbstständigkeitserklärung **171**

Lebenslauf **173**

Erklärung zu den Eigenanteilen und Bestätigung des Einverständnisses der Koautoren **175**

Einverständniserklärung der Betreuerin **177**

List of Figures

1.1 The development of the 2003 European drought is illustrated by depicting the drought extent for February, May, and August of 2003. A drought is occurring when the soil moisture index is below 0.2 indicating a low fractional soil moisture value that is observed less than 20% of the time. These fields are obtained by driving the mesoscale Hydrologic Model (mHM) with the observation-based E-OBS dataset. 36

1.2 Schematic representation of a drought prediction system. This Figure is divided vertically into an observation-based part (left of vertical dashed line) and a forecasting part (right of vertical dashed line). The observation-based part is comprised of: first, an observation-based meteorological dataset; second, a hydrologic model that is driven with these observations. Third, the simulated soil moisture is used for reconstruction during historic periods and to provide initial hydrologic conditions (IHCs) for the forecasts. The forecasting part (i.e., the right hand side of this figure) consists of: first, a meteorological forecast obtained from a climate model (or numerical weather prediction model); second, a downscaling of climate model outputs that requires also the meteorological observations; third, a hydrologic model of the land surface that is driven with the downscaled meteorological forecasts to provide a soil moisture forecast. This work is mainly focusing on the evaluation of climate model outputs with respect to extreme meteorological indices (Chapter 2) and the temporal downscaling of meteorological variables (i.e., disaggregation method introduced in Chapter 3). The evaluation of the drought prediction system is then presented in Chapter 4. 37

1.3 Standardized soil moisture anomalies for the period from January to December 2003. The red line depicts the reference anomaly obtained by the E-OBS dataset. The blue and green lines show the soil moisture anomalies of NMME-based forecasts with one month lead time. 40

1.4 Outputs of climate models (CM) are typically available at lower spatial and temporal resolutions than the input required for hydrologic models (HM). Temporal disaggregation schemes such as those presented in Chapter 3 increase the temporal resolution of CM to that of HM (horizontal arrows). Spatial downscaling such as those discussed in Section 5.2 increase the spatial resolution (vertical arrows). The thickness of the arrows corresponds to how simple a particular step is in comparison to other ones. 42

2.1 Percentage of total variance explained by a given number of principal components for the aggregated REGNIE product (red lines) and the EOBS data set (blue lines) for two periods: summer (red lines) and winter (blue lines). 60

2.2 Spatial average of relative bias B for every index and RCM. Panels a) and b) depict the performance of precipitation and temperature indices, respectively. The statistical summary of theses indices is shown in Table 2.3. 61

2.3 Spatial Pattern for each temperature index (rows). The spatial pattern for the observations is shown in the first column. The units for each index are encompassed in Table 2.2. The second, third, and fourth column show the relative bias in percent for the worst performing model for each index, the best performing model for each index, and the ensemble average of all models. The worst- and best-performing RCMs are assessed with respect to the interpolated observational data set. 64

2.4 Analogous to Fig. 2.3, only for precipitation indices. The worst- and best-performing RCMs are assessed with respect to the REGNIE observational data set. 66

2.5 Spatially averaged weights for each index derived by the "Reliability Ensemble Averaging" (REA/method 5) for each model. Red and blue markers indicate temperature and precipitation indices, respectively. 67

2.6 Final score S (panel a) and rejection rate $\overline{R}$ (panel b) for single models and selected subensembles identified by different methods. Different search methods are indicated by: O - method 1, S - method 2, W - method 3, F - method 4, and B - method 5. For clarity, the x-axis position of the markers are varied. 68

2.7 Top panel a): Variability of the rejection rate $\overline{R}$ for all ensembles of different sizes is shown with box-plots. The box-plots range from minimum to maximum $\overline{R}$ value with 50% of the $\overline{R}$ values between the 25th and 75th percentile in the box. The median is marked as red bar. Furthermore, the performance of the selection methods is shown. Yellow dots represent method 4, red circles method 5, black dots method 2, and black stars method 3. The $\overline{R}$ of the subensemble selected by method 1 is marked by the lower bar. For clarity, a logarithmic y-axis is used. Bottom panel b): The spatial performance for ensembles of nine models selected according to method 2, 3, 4, and 5 (i.e. S9, W9, F9, and B9, respectively) is shown in the first row. The spatial performance for the ensemble of nine models selected by method 1 (O9), the full ensemble (13), and two single models (the best - G and the second best - E) is shown in the second row. The rejection rate was estimated at the 5% significance level. . . . 70

2.8 Analogous to Fig. 2.7 without considering model G. Bottom panel b): The spatial performance for ensembles of eight models selected according to method 2, 3, 4, and 5 (i.e. S8, W8, F8, and B8, respectively) is shown in the first row. The spatial performance for the ensemble of eight models selected by method 1 (O8), the full ensemble (12), and two single models (the second best - E and the third best - L) is shown in the second row. 72

3.1 Diagram for disaggregation of a precipitation time series at scale K to a precipitation time series at scale k. 83

3.2 Preprocessing required for the proposed method as well as algorithmic steps required for the estimation and the generation part of the proposed method. The estimation steps are performed for every temporal scale $K = 1, 2, 4, \ldots, 32$. 84

3.3 Map of long term annual precipitation [mm] across Germany including the positions of six locations representing different precipitation regimes: (1) Schleswig, (2) Dortmund, (3) Brocken, (4) Berlin, (5) Feldberg, and (6) Munich. 88

3.4 Cumulative distribution function (cdf) $\mathcal{F}$ of the weights w between the two-day and daily resolution for (a) location (4) Berlin and (b) location (5) Feldberg estimated during the calibration period of 1950-1990. Colored lines indicate the cdfs for different intensity classes (Table 3.1) ranging from low (red) to high intensities (blue). Gray lines show the cdfs assuming only one intensity class. 89

3.5 Quantile-Quantile plot between the cdfs $\mathcal{F}$ for the partitioning of two-day precipitation estimated during the calibration and evaluation period at six contrasting locations. The cdfs are depicted for three selected intensity classes ranging from low (red lines) to high ones (blue lines). 90

3.6 One exemplary disaggregation cascade for a single 32 day field. The unit throughout the panels is [mm]. The top left plot is an observed 32 day value. It is depicted in each row how a low resolution value (left plot) is stochastically disaggregated to two higher resolution values (right plots). 92

3.7 Evaluation of deciles of the precipitation distribution function at every grid cell. (a) The rejection rate (RR) and (b) the mean absolute relative bias (MARB) over all deciles $D_1, \ldots, D_9$ during the calibration period. Second row: same as the first row, but for the evaluation period. 94

3.8 Panel (a) shows the differences in observed precipitation distribution functions between calibration and evaluation periods. For comparison, panel (b) shows the difference between the simulation and observation during the evaluation period (same as Figure 3.7d). The differences are calculated as the mean absolute relative bias (MARB) over all deciles $D_1, \ldots, D_9$ of the respective distribution functions. . . . 95

3.9 Exceedance probabilities of daily precipitation depicted for six selected locations (Figure 3.3) during the evaluation period 1991-2010. Dots mark the observed values at exceedance probabilities of $0.1\%, 1\%, 5\%, 10\%, 20\%, 30\%, \ldots, 100\%$. Solid lines depict the median of simulations and dashed lines the 95% confidence interval. The horizontal dashed line shows the 50% exceedance probability. 96

3.10 Probabilities of dry days for (a,d) the observations and (b,e) the ensemble median of simulations during the calibration (1950-1990) and the evaluation period (1991-2010). The relative bias between the ensemble median of simulations and the observations is shown in panels (c) and (f) during the calibration and evaluation period, respectively. 97

3.11 Median of absolute relative bias (ARB) between the simulated daily ensemble and the observed daily values for the extreme indices (a) $r95p$ and (b) CDT at every cell during the evaluation period. $r95p$ denotes the sum of rainfall above the 95^{th} percentile and CDT denotes the maximum number of consecutive dry time steps. Boxplots show the ensemble variability of relative bias for (c) $r95p$ and (d) CDT with respect to the observations (dashed line) at six selected locations (Figure 3.3). The median of the ensemble is represented by a bold black line. The box ranges from the 25^{th} to the 75^{th} percentile. The whiskers show the range of the 95% confidence interval and the asterisks mark the minimum and maximum value. 98

3.12 Pearson cross-correlation coefficients ρ between specific locations and all other locations averaged over distance in bins of 5 km width during the evaluation period. The Pearson cross correlation ρ is shown for the observations (solid lines) and simulations (dashed lines). Red lines have been calculated for the weekly time scale (8 d). For the daily scale, blue lines have been estimated using the whole data set for all months, brown lines for summer months (MJJASO), and green lines for during winter months (NDJFMA). Additionally, the 95% confidence interval is depicted as bands. The Pearson cross correlation is shown for six selected locations (Figure 3.3). 100

3.13 Boxplots of relative bias between the median of simulations and observations over all cells during the evaluation period for a specific temporal resolution and statistic. The presented statistics are (a) all deciles, (b) the index $r95p$, (c) the index CDT, and (d) the Pearson cross correlation coefficient ρ. The median of relative bias is marked as black line. The box ranges from the 25^{th} to the 75^{th} percentile. The whiskers show the spread of the 95% confidence interval and the asterisks mark the minimum and maximum value. 102

3.14 Autocorrelation γ for daily values up to ten lag days at six locations (Figure 3.3) during the evaluation period. Solid lines depict the autocorrelation for the observations. Dashed lines show the median autocorrelation for the simulations and the band quantifies the 95% confidence interval of simulations. 103

4.1 The skill to reproduce reference SMI is illustrated in terms of the Pearson correlation coefficient R between the forecasted and reference SMI for lead times of one, three, and six months. The skill of the NMME ensemble is depicted for two averaging schemes: $\overline{\text{SMI}}$ and $\text{SMI}(\overline{\text{SM}})$ in panels a-c and d-f, respectively. In the panels g-i, the skill of the ESP approach is shown. The persistence of reference SMI (estimated as Pearson auto-correlation) is displayed in the panels j-l. The spatial average of the corresponding R is depicted in the upper right corner of each panel. 116

4.2 For a given grid cell (located in Central France at 47.19° N, 3.21° E), the exemplary time series of SM and SMI are depicted in panels a and b, respectively. In both panels, the blue line delineates the dynamics of the reference dataset and the gray band shows the uncertainty obtained from the 24 ensemble members of the CFSv2 forecasts at two months lead time. The gray dashed line in the top panel a denotes the average of the CFSv2 SM ensemble. The gray and black dashed lines in the bottom panel b denote the SMI ensemble derived by the $\overline{\mathrm{SMI}}$ and $\mathrm{SMI}(\overline{\mathrm{SM}})$ averaging scheme, respectively. The thin horizontal dashed line illustrating the drought threshold 0.2 is displayed for clarity. 117

4.3 In the top panel a, the overall Pearson correlation and ETS estimates are shown for the NMME subensembles (red bars), single models (blue bars), and ESP (gray bar). These estimates are averaged over space and lead times to meet the requirements of a seamless prediction system. The SMI of NMME subensembles is obtained by the $\mathrm{SMI}(\overline{\mathrm{SM}})$ averaging scheme. In the bottom panel b, the single models contained within a selected subensemble for Pearson correlation and ETS are depicted by blue boxes. 119

4.4 Spatial distribution of ETS at one, three, and six month lead time is displayed for the full NMME ensemble (NMME8) in panels a-c, NMME subensemble containing four models (NMME4) in panels d-f, and ESP in panels g-i. The NMME based forecasts are obtained by the $\mathrm{SMI}(\overline{\mathrm{SM}})$ averaging scheme. The corresponding spatial averages of ETS are denoted in the upper right corner of every panel. 120

4.5 The top panel depicts the fraction of area under drought based on the reference SMI dataset. The thin horizontal dashed line is added for clarity displaying the threshold for droughts covering more than 20% of the European domain. Panels b-d illustrate the temporal variability of ETS for the ESP (blue lines) and the full NMME ensemble (red lines) based SM drought forecasts. The NMME based forecasts are obtained by the $\mathrm{SMI}(\overline{\mathrm{SM}})$ averaging scheme. Additionally, the 95% confidence interval for the single ESP and NMME ensemble members is depicted as light red and blue bands, respectively. Ticks mark the end of the respective year. The scale of y-axis are different for each panel for clarity. 122

4.6 Probability density function for drought severity and drought area is illustrated for different lead times for forecasts obtained by single NMME and ESP ensemble members. The performance for all NMME models is shown in panels a-c and only for four NMME models in panels d-f. The performance for ESP based forecasts is displayed in panels g-i. The area containing 90% of the density for both characteristics is depicted in each panel as red and blue regions for drought development and recovery, respectively. Additionally, the spread for each characteristic is shown as box plots for the different drought phases (95% confidence interval as thin lines, the spread between the 25th and 75th quantiles as thick lines, and the median is located at the intersection). 124

5.1 Two dimensional histogram of monthly precipitation distribution functions at a 8 km (x-axis) and 4 km (y-axis) spatial resolution. The bin width of the histogram is 1 percent in both distribution functions. The marginal distributions have been estimated as empirical distribution functions of the same dataset investigated in Chapter 3 (i.e., gridded precipitation dataset with a 4×4 km^2 spatial resolution over entire Germany available for the period from 1961 to 2010). This figure is identical to the empirical copula for the two quantities. 131

5.2 One example for a copula downscaling of WRF precipitation from a 12×12 km^2 to a 4×4 km^2 resolution. This downscaling is conducted for the monthly precipitation amount of August 1992. The WRF data has been kindly provided by Dr. Kirsten Warrach-Sagi (University of Hohenheim). 132

5.3 Relative biases for extreme meteorologic indices investigated in this work. The two left plots depict the absolute relative biases (ARB) for the precipitation amount above the 95th percentile ($r95p$) and the maximum number of consecutive dry time steps (CDT) of the ensemble of disaggregated monthly precipitation fields analyzed in Chapter 3 (adapted from Figure 3.11a and 3.11b). The right plots show the same indices ($r95p$ separated for summer $r95psum$ and winter $r95pwin$ and CDD which is identical to CDT for daily time steps) and the total annual precipitation $RAnn$ for the worst and best performing regional climate model and the full ensemble mean of 13 regional climate models investigated in Chapter 2 (adapted from Figure 2.4). 135

C.1 Frobenius norm of the difference matrix between observed and simulated cross-covariance matrix for different parameter settings. The distance thresholds D are 100 km, 200 km, and 300 km for the first, second, and third column, respectively. The number of neighboring cells L are 100, 200, and 300 for the first, second, and third row, respectively. The selected number of anchor cells A are 0, 20, 40, 60, and 80. These are displayed on the x-Axis of each plot. The 5.0×10^8 line (red dashed line) is displayed to ease the comparison among different plots. 149

C.2 Spatial distribution of k-means cluster with $k = 4$ of (a) precipitation distribution functions and (b) cdfs $\mathcal{F}$ between two day and daily scale for one intensity class over Germany. 151

List of Tables

2.1 Regional Climate Models incorporated in this study from the ENSEMBLES project (van der Linden and Mitchell, 2009) 53

2.2 Temperature (upper half) and precipitation (lower half) indices incorporated in this study. 54

2.3 Spatial average of long term values for each Temperature and Precipitation index for the observations, all single models, and six ensembles. These are the full ensemble (13), the optimal ensemble of size nine (O9), the score based ensemble of size nine (S9), the weighted ensemble average of size nine (W9), the forward selected ensemble of size nine (F9), and the backward selected ensemble of size nine (B9). * indicates that the null hypothesis is rejected in 50% of the cell with the 5% significance level, where as ** and *** indicate that 50% of the cells are significant at the 1% and 0.1% significance level, respectively. Additionally the aggregated scores for Precipitation and Temperature as well as the ranking of these are shown. 62

3.1 Upper thresholds [mm] for precipitation intensity classes c estimated for the observed data set for each temporal resolution during the calibration period 1950-1990. The lower threshold of a particular class is the upper threshold of the preceding class. The lower threshold of the first class is zero. 89

3.2 Percentage of cells exhibiting a given level of (a) rejection rate (RR) and (b) mean absolute relative bias (MARB) with respect to all deciles. Tables (c) and (d) are the same as (a) and (b), but for high percentiles, which are the 95^{th}, 99^{th}, and 99.9^{th} percentile. 93

4.1 Climate Forecasting models used in this study, Institute they are developed at, and ensemble members available (see Kirtman et al. (2014) for details). 112

Danksagung

Wissenschaftliches Arbeiten ist keine Einzelarbeit und auch keine Einzelleistung, sondern hat den regen Austausch von Ideen, Meinungen und Ratschlägen als notwendige Bedingung. Ich hatte in meiner Zeit als Doktorand das Glück mit Menschen zusammenarbeiten zu dürfen, die stetig bemüht waren mir den Weg zum wissenschaftlichen Arbeiten durch unermüdliche Diskussionen und Kritik zu weisen. Diesen gilt im Folgenden mein Dank.

Als Erstes möchte ich mich bei *Prof. Dr. Sabine Attinger* dafür bedanken, dass sie mir die Gelegenheit gab diese Dissertation anzufertigen.

Mein besonderer Dank gilt *Dr. Luis E. Samaniego* für die Betreuung dieser Arbeit, die mir entgegengebrachte Begeisterung für mein Thema, das präzise Korrekturlesen meiner Manuskripte und insbesondere für das Augenmerk auf Stringenz, die nicht immer leicht zu befriedigen war, nichtsdestotrotz aber notwendig ist. Des Weiteren möchte ich mich auch für die Unterstützung und Ermunterung zur Teilnahme an internationalen Konferenzen und die dabei gewonnenen Kontakte bedanken.

Dr. Juliane Mai danke ich für das Korrekturlesen meiner Manuskripte und insbesondere die Verbesserungsvorschläge, die mir eine wahre Anleitung zum Verfassen wissenschaftlicher Artikel bedeuteten.

Bei *Dr. Matthias Cuntz* möchte ich mich nicht nur für die zahlreichen Diskussionen jedweder Couleur bedanken, sondern vor allem auch für all die Ratschläge im Umgang mit dem Computer (Emacs, Bash, Python, Fortran, Makefile, netCDF, cdo, nco,...). Ohne diese Hilfe hätte die Anzahl der Tastenanschläge nicht gereicht, um diese Arbeit anzufertigen.

Matthias Zink und *Dr. Rohini Kumar* danke ich für die zahlreichen lebhaften Diskussionen zur hydrologischen Modellierung, welche mir dieses für mich neue Fachgebiet näher gebracht haben. Auch gilt mein Dank allen Kollegen in der Hydrosystemmodellierung für die gute Arbeitsatmosphäre, in der diese Arbeit erst entstehen konnte.

Bei *Prof. Dr. András Bárdossy* bedanke ich mich für die Diskussionen und Ideen, welche diese Arbeit maßgeblich beeinflusst haben.

Mein besonderer Dank gilt meiner Familie. Insbesondere meinen Großeltern und meiner Mutter für das in mich gesetzte bedingungslose Vertrauen. Ich danke Dir, Jule, für Deine Unterstützung und unser wundervolles Leben.

Übersicht

Diese kumulative Dissertation besteht aus drei Publikationen. Kapitel zwei ist inhaltlich identisch zu der Publikation:

1.) Stephan Thober and Luis Samaniego, “Robust Ensemble Selection by Multivariate Evaluation of Extreme Precipitation and Temperature Characteristics”, Journal of Geophysical Research: Atmospheres, 119(2):594–613, 2014, ISSN 2169-8996, doi:10.1002/2013JD020505.

Kapitel drei ist inhaltlich identisch zu der Publikation:

2.) Stephan Thober, Juliane Mai, Matthias Zink, and Luis Samaniego, “Stochastic Temporal Disaggregation of Monthly Precipitation for Regional Gridded Data Sets”, Water Resources Research, 50(11):8714–8735, 2014, ISSN 1944-7973, doi:10.1002/2014WR015930.

Kapitel vier ist inhaltlich identisch zu dem Manuskript:

3.) Stephan Thober, Rohini Kumar, Justin Sheffield, Juliane Mai, David Schäfer, und Luis Samaniego, “On the Capability of the North American Multi-Model Ensemble for Seasonal Soil Moisture Drought Prediction over Europe”, am 7. April 2015 beim Journal of Hydrometeorology eingereicht. Am 20. Mai 2015 wurde das Manuskript zur Pubklikation mit wenigen Änderungen akzeptiert, welches die Einreichung belegt. Der Entscheidungsbrief des Editors ist untenstehend zu finden.

> Dear Dr. Samaniego,
>
> I am now in receipt of all reviews of your manuscript "On the Capability of the North American Multi-Model Ensemble for Seasonal Soil Moisture Drought Prediction over Europe". I am happy to report that the reviewers unanimously recommended an editorial decision of Minor Revision, and on reading the paper, I agree with their assessment. Copies of the reviews are enclosed below or attached as files to this letter.
>
> We invite you to submit a revised paper by Jul 18, 2015. If you anticipate problems meeting this deadline, please contact me as soon as possible at Wood.JHM@ametsoc.org.
>
> Along with your revision, please upload a point-by-point response that satisfactorily addresses the concerns and suggestions of each reviewer. Should you disagree with any of the proposed revisions, you

will have the opportunity to explain your rationale in your response. In addition, I ask that you give careful consideration to the comment (of Reviewers 1 and 3) that the predictability of SM varies seasonally. I think it would strengthen the paper to either discuss thoughtfully how your results (such as the strength and selection of the combined ensembles) would be influenced by this factor, or even add analysis to provide insight on this point.

Please visit http://www.ametsoc.org/PUBSrevisions to view the AMS Guidelines for Revisions. Be sure to meet all recommendations on the guidelines for quickest processing of the revised manuscript.

When you are ready to submit your revision, go to http://amsjhm.edmgr.com/ and log in as an Author. Click on the menu item labeled "Submissions Needing Revision" and follow the directions for submitting the file.

Thank you for your patience while a decision was reached.

Sincerely,

Dr. Andrew Wood
Editor
Journal of Hydrometeorology

Gesamtzusammenfassung

Wasser ist ein essentielles Element für alles bekannte Leben auf der Erde und befindet sich in einem ständigen Kreislauf zwischen den verschiedenen Teilen des Erdsystems, wie zum Beispiel dem Ozean, dem Land, dem Untergrund und der Atmosphäre. Der Mensch beeinflusst den natürlichen Wasserkreislauf auf verschiedenste Art und Weise, um seine Lebensgrundlage zu sichern. Es werden beispielsweise Deiche errichtet zum Schutz von Eigentum vor Hochwasser und Dämme werden gebaut zur Stromerzeugung durch Wasserkraft und zum Anlegen von Reservoiren für die Trinkwasserversorgung.

Hydrologische Extremereignisse stellen jedoch weiterhin eine Bedrohung für die Lebensgrundlage des Menschen dar. Hochwasser und Dürren sind die beiden hydrologischen Extremereignisse, welche auf dem gesamten Globus zu mehr Todesfällen führen, die höchsten wirtschaftlichen Verluste verursachen, und mehr Menschen betreffen als jede andere Naturkatastrophe. Diese beiden Phänomene sind sehr unterschiedlich in ihrer Natur. Ein Hochwasser ist ein sich schnell entwickelndes Ereignis, welches innerhalb von Tagen bis Wochen in einem Flusseinzugsgebiet auftritt. Im Gegensatz dazu ist eine Dürre ein langsames Ereignis, das sich über Monate und Jahreszeiten entwickelt, Jahre überdauert und ganze Regionen, Länder sowie Kontinente betrifft. Dürren sind definiert als ein Wasserdefizit im Vergleich zu normalen Bedingungen und treten in allen Kompartimenten des Wasserkreislaufs auf, wie zum Beispiel in der Atmosphäre (meteorologische Dürre), in Flüssen und im Grundwasser (hydrologische Dürre) und in Böden (landwirtschaftliche Dürre).

Landwirtschaftliche Dürren führen zu hohen sozialen und wirtschaftlichen Schäden sowie humanitären Krisen (z.B. Hungersnöte), weil sie das Potenzial haben Ernteerträge zu verringern. Die Auswirkungen dieser Extremereignisse können mit der Hilfe von einem saisonalen Vorhersagesystem, welches Dürren mehrere Monate im Voraus prognostiziert, abgeschwächt werden. Diese Arbeit ist somit der Entwicklung eines saisonalen Vorhersagesystems für landwirtschaftliche Dürren innerhalb Europas gewidmet. Dieses Vorhersagesystem basiert auf meteorologischen Prognosen des nordamerikanischen Multi-Modell Ensembles (NMME), welche verwendet werden um das mesoskalige Hydrologische Modell (mHM) anzutreiben. Der NMME Datensatz wird in dieser Arbeit betrachtet, da er ein relativ großes Ensemble beinhaltet und ein einfacher Datenzugriff gewährleistet ist. Das in dieser Arbeit entwickelte Vorhersagesystem ist jedoch flexibel genug um auch andere Datensätze, wie beispielsweise TIGGE-LAM vom Europäischen Zentrum für mittelfristige Wettervorhersagen (ECMWF), zu verwenden.

Der NMME Datensatz stellt monatliche Niederschlags- und Temperaturprognosen bereit. Tageswerte sind zwar auch verfügbar, diese erfordern jedoch wesentlich höhere Speicherkapazitäten und sind daher für Anwender oft von geringem Interesse. Obwohl die acht Modelle in diesem Ensemble sich in ihrer Modellstruktur unterscheiden (d.h. in der Auswahl und der Parametrisierung der implementierten Prozesse), bilden sie alle den Wasser- und Energiekreislauf der Erde als dynamisches System ab. Das mesoskalige Hydrologische Modell (mHM) wird in dieser Arbeit benutzt, um den Wasserkreislauf auf der Landoberfläche zu berechnen, da dieses räumlich verteilte Modell dafür entwickelt und kalibriert wurde die Wasserbilanz an der Mündung eines Einzugsgebietes zu schließen. Dies wird durch das zuverlässige Aufteilen von Niederschlag in Evapotranspiration zur Atmosphäre und Oberflächenabfluss in die Ozeane erreicht. Diese Partitionierung wird durch die Modellierung der Bodenfeuchte realisiert. Die erhaltenen Bodenfeuchtefelder von mHM werden dann zu einem Quantil basierten Index transformiert, um eine Dürreanalyse durchzuführen.

Die Güte der Dürrevorhersage ist von mehreren Faktoren abhängig. Hauptbestandteile des saisonalen Vorhersagesystems sind u.a. die meteorologischen Prognosen, welche von den NMME Modellen bereitgestellt wurden. Ein Standardansatz in der Dürrevorhersage ist es entweder die Prognosen der einzelnen Modelle oder die des gesamten Ensemblemittels auszuwerten (d.h. das Mittel über die Prognosen aller einzelnen Modelle). In dieser Arbeit werden Methoden untersucht, die über einen solchen Ansatz hinausgehen, um die Fähigkeit von Multi-Modell Ensembles zur Abbildung von Extremindizes, welche für die Hochwasser- und Dürremodellierung relevant sind, zu verbessern. Es wird in dieser Arbeit die Hypothese aufgestellt, dass eine Untergruppe von Modellen (d.h. ein Subensemble) eine bessere Vorhersagegüte liefert als einzelne Modelle und das gesamte Ensemblemittel. Verschiedene neu entwickelte Methoden werden untersucht, um Subensembles mit der besten Performance effizient auszuwählen ohne alle möglichen Kombinationen von Modellen zu betrachten.

Auswahlalgorithmen für Subensembles werden für eines der neuesten Ensembles von regionalen Klimamodellen getestet, welches vom ENSEMBLES Projekt zur Verfügung gestellt wurde. Die räumliche und zeitliche Variabilität von elf extremen Niederschlags- und Temperaturindizes, welche mit Hochwasser und Dürren in Zusammenhang stehen, wird betrachtet, um die Güte der Modellsimulationen zu bewerten. Die geschätzten Indizes aus diesen Modellsimulationen werden mit jenen aus Beobachtungsdaten über Deutschland in der Zeit von 1961 bis 2000 verglichen. Die Güte der einzelnen Modelle und der Subensembles wird mit Hilfe einer Ablehnungsrate quantifiziert, welche die statistische Signifikanz der erhaltenen Abweichungen auf Basis des Wilcoxon-Rangsummentests auswertet. Die regionalen Klimamodelle weisen erwartungsgemäß einen geringeren Bias für Temperaturindizes im Vergleich zu Niederschlagsindizes auf. Die Schwankungsbreite des Bias für extreme Niederschlagsindizes ist in der Regel $\pm 20\%$. Des Weiteren über- und unterschätzen diese regionalen Klimamodelle bestimmte Indizes systematisch. Beispielsweise wird der jährliche Gesamtniederschlag überschätzt, während die maximale Anzahl der aufeinander folgenden trockenen Tage unterschätzt wird.

Für die regionalen Klimamodelle des ENSEMBLES Projekts zeigte das beste Subensemble kleinere und weniger signifikante Bias als alle einzelnen Modelle und das gesamte Ensemblemittel in verschiedenen Teilen Deutschlands. Eines der vorgeschlagenen Verfahren, der Rückwärtseliminationsalgorithmus, ist in der Lage dieses beste Subensemble effizient zu finden ohne alle möglichen Subensembles evaluieren zu müssen. Das beste Subensemble enthält nur 6 der 13 betrachteten regionalen Klimamodelle. Die Verwendung dieses Subensembles statt des gesamten Ensemblemittels kann somit zu abweichenden Schlussfolgerungen in Klimafolgenabschätzungen führen.

Die räumliche und zeitliche Auflösung eines Klimamodells ist in der Regel niedriger als die Auflösung eines hydrologischen Modells. Für das hier entwickelte Dürrevorhersagesystem entspricht die zeitliche Auflösung der meteorologischen NMME Prognosen Monatswerten, während die hydrologische Modellierung Tages- oder Stundenwerte als Eingangsdaten benötigt. Ein neu entwickeltes stochastisches Verfahren wird benutzt um die monatlichen Niederschlagsfelder in tägliche zu disaggregieren. Diese Methode behandelt zunächst nur Niederschlag, da dieser eine schiefe Verteilungsfunktion und eine komplexe zeitliche und räumliche Kovarianzstruktur aufweist. Dieser Ansatz basiert auf einer multiplikativen Kaskade, welche einen Niederschlagswert von einer niedrigen zeitlichen Auflösung in zwei Werte auf einer höheren Auflösung partitioniert (z.B. einen zweiwöchentlichen Wert in zwei wöchentliche) und dabei die jeweilige räumliche Kovarianzstruktur auf der entsprechenden zeitlichen Auflösung erhält. Dieses Kaskadenverfahren beginnt mit Monatswerten und stoppt sobald Tageswerte erhalten werden. Die Partitionierung erfolgt indem Gewichte (d.h. multiplikative Faktoren) entsprechend einer Verteilungsfunktion gesampelt werden. Diese Verteilungsfunktionen werden für unterschiedliche Niederschlagsintensitäten aus Beobachtungsdaten bestimmt und es wird gezeigt, dass sich diese nicht wesentlich im Laufe der Zeit verändern. Dies macht diese Methode auch für zukünftige Perioden anwendbar. In dieser Arbeit wird die Hypothese aufgestellt, dass die disaggregierten täglichen Felder statistisch äquivalent zu den beobachteten sind.

Dieses Disaggregierungsverfahren wurde vor allem entwickelt um die zeitliche Auflösung eines Datensatzes zu erhöhen. Dabei erzeugt es aber auch Felder mit einer konsistenten räumlichen Kovarianzstruktur. Dieses wird durch eine neu entwickelte sequentielle Sampling-Technik, welche “Anchor Sampling” genannt wird, erreicht. Diese Methode ist auf Rastergittern von beliebiger Auflösung und Größe anwendbar. Das “Anchor Sampling” erfordert keine Annahme über die räumliche Isotropie von Niederschlagsfeldern. Dies ist von elementarer Bedeutung, weil orographische Effekte eine nicht isotrope räumliche Kovarianzstruktur erzeugen. Diese Methode kann auch für andere stochastische Methoden von Vorteil sein, welche auf die Generierung von räumlichen Feldern mit einer konsistenten Kovarianzstruktur angewiesen sind, wie zum Beispiel räumliche Skalierungsverfahren.

Das stochastische Disaggregierungsverfahren wird für hochaufgelöste Niederschlagsdaten (d.h. $4 \times 4\,\mathrm{km}^2$ Raster) über Deutschland ($\approx 357\,000\,\mathrm{km}^2$) in der Zeit von 1950 bis 2010 getestet. Die Auswertung des Verfahrens erfolgt für einen Kalibrierungszeitraum (1950-1990) und einen Validierungszeitraum (1991-2010), um die Übertragbarkeit auf zukünftige Perioden zu testen. Die Ergebnisse zeigen, dass ortsabhängige Verteilungsfunktionen vom Niederschlag mit Abweichungen von weniger als 5% während des Validierungszeitraumes reproduziert werden. Zusätzlich sind extreme Niederschlagsindizes mit einem Bias von weniger als 10% abgebildet. Diese Fehler sind kleiner als jene, die für die regionalen Klimamodelle des ENSEMBLES Projekts beobachtet wurden (für die gleichen Indizes). Dies hebt den Mehrwert des vorgeschlagenen Disaggregierungsverfahrens hervor. Es ist erwähnenswert, dass die zeitliche Autokovarianz in den disaggregierten Niederschlagsfeldern vergleichbar ist mit jener der beobachteten Felder. Dies wird erreicht ohne eine explizite Annahme über die Autokovarianz der disaggregierten Felder zu formulieren und kann daher als eine Eigenschaft betrachtet werden, die durch die Kaskadenmodellstruktur erzeugt wird.

Das saisonale Dürrevorhersagesystem enthält dann den Rückwärtseliminationsalgorithmus, um das Subensemble mit der höchsten Vorhersagegüte des NMME Datensatzes zu identifizieren. Es nutzt auch die neu entwickelte stochastische Disaggregierungsmethode, um die monatlichen Niederschlagsprognosen in tägliche zu überführen. Andere für mHM benötigte Eingangsvariablen, wie zum Beispiel Temperatur, werden durch eine Skalierung von historischen Feldern erhalten. Diese haben nach der Skalierung den gleichen monatlichen Wert wie die Vorhersagen der NMME Modelle. Die von dem saisonalen Dürrevorhersagesystem erhaltenen Prognosen werden mit denen einer einfachen statistischen Methode, die auf dem Ensemble Streamflow Prediction (ESP) Ansatz basiert, verglichen. Der ESP Ansatz benutzt meteorologische Beobachtungen aus der Vergangenheit zur Erstellung eines Vorhersageensembles für den Antrieb von mHM. Die Vorhersagen dieses Verfahrens basieren somit nur auf klimatologischen Informationen und enthalten keine Kenntnis über die tatsächliche Entwicklung des Erdsystems. Der Bewertungszeitraum des Dürrevorhersagesystems reicht von 1983 bis 2009 mit monatlichen Vorhersagen beginnend am Anfang eines jeden Monats und Vorhersagezeiten von bis zu 6 Monaten. Das Simulationsgebiet umfasst große Teile Kontinentaleuropas. Der Equitable Threat Score (ETS) wird in dieser Arbeit benutzt, um die Vorhersagegüte zu quantifizieren. Dieser fasst die Trefferquote und die Falsch-Positiv-Rate in einem Bewertungsmaß zusammen. Der frei verfügbare E-OBS Datensatz wird genutzt, um Referenzwerte für die Bodenfeuchte zu erhalten. Es wird in dieser Arbeit die Hypothese aufgestellt, dass die NMME basierten Prognosen einen höheren ETS aufweisen als die ESP basierten.

Die auf dem NMME basierenden Bodenfeuchteprognosen zeigen einen ETS auf, der bei einer sechsmonatigen Vorhersagezeit im Durchschnitt 69% höher ist als der des ESP Ansatzes. Die Ergebnisse zeigen, dass es eine substantielle räumliche und zeitliche Variabilität in der Güte der Dürrevorhersagen von bis zu 40% gibt. Die Vorhersagegüte ist beispielsweise für beide Prognoseverfahren (d.h. NMME und ESP basierte Prognosen) höher in Regionen in denen die Referenzbodenfeuchte selbst eine hohe Persistenz aufweist. Die perfekte Kenntnis der initialen hydrologischen Bedingungen führt dann zu einer hohen Vorhersagegüte in diesen Regionen. Niedrige Vorhersagegüten sind vor allem in Zeiten zu beobachten wenn keine Dürre vorliegt. Dies bedeutet, dass beide Vorhersagemethoden das Ausmaß von Trockenheiten in diesen Zeiträumen überschätzen. Die Güte von Dürrevorhersagen ist somit sehr stark von dem Dürrezustand zum Beginn der Prognose abhängig, welcher in den initialen Bedingungen enthalten ist.

Unter den auf dem NMME basierenden Prognosen besitzt das gesamte Ensemblemittel einen höheren ETS als jedes identifizierte Subensemble und das bestmögliche Einzelmodell. Die Vorhersagegüte der Subensembles, die mehr als drei Modelle enthalten, ist jedoch nur 1% geringer als die von dem gesamten Ensemblemittel. Die Anzahl der Modellläufe dieser Subensembles beträgt jedoch nur 60% der des gesamten Ensembles. Diese Subensembles könnten somit für operative saisonale Dürrevorhersage und Anwender, welchen nur begrenzte Rechenkapazitäten zur Verfügung stehen, nützlich sein. Im Allgemeinen sollte das Dürrevorhersagesystem, welches auf meteorologischen Vorhersagen des NMME basiert, für den operationellen Betrieb in Betracht gezogen werden, weil es eine höhere Vorhersagegüte als klimatologische Vorhersagen über ganz Europa und alle Vorhersagezeiten aufweist.

Die in dieser Arbeit vorgestellten Methoden sind nicht auf die hier betrachtete Anwendung der saisonalen Dürrevorhersage beschränkt. Stattdessen sind sie allgemein und auch für andere Anwendungen genauso nützlich. Beispielsweise kann das Rückwärtseliminationsverfahren auch auf andere Ensembles und Gütekriterien angewendet werden. Basierend auf den Ergebnissen dieser Arbeit wird erwartet, dass diese Methode die Nützlichkeit von Modell Ensembles auf zwei Arten verbessern kann. Erstens kann ein Subensemble eine höhere Performance erzielen als das gesamte Ensemblemittel und zweitens kann ein Subensemble eine nahezu gleichwertige Performance erreichen wie das gesamte Ensemblemittel, aber mit einem niedrigeren Rechenaufwand.

Das in dieser Arbeit vorgestellte "Anchor Sampling" ermöglicht die Generierung räumlicher Felder normalverteilter Zufallszahlen mit einer vordefinierten räumlichen Kovarianzstruktur. Diese Methode wurde nur für die zeitliche Disaggregierung von monatlichen Niederschlagswerten in tägliche genutzt. Folgestudien sollten den "Anchor Sampling" Ansatz auch für andere Anwendungen wie zum Beispiel das räumliche Skalieren von Klimamodelldaten verwenden.

Abstract

Water is a vital element for all known life forms and is constantly cycling through different compartments of the Earth system such as the ocean, atmosphere, land surface, and subsurface. Humans modify the water cycle on land in various ways to ensure their livelihood. For instance, levees are constructed for protecting property from floods, dams are built to generate hydropower and to create reservoirs for drinking water supply, and groundwater is pumped for freshwater supply.

Hydrologic extremes, however, keep on threatening the livelihood of human societies. The two most devastating hydrologic extremes are floods and droughts resulting in more deaths, yielding the highest economic losses, and affecting more people than any other natural hazard on a global scale. These two phenomena are very different in their nature. A flood is a fast evolving event that is spanning over days to weeks and is occurring at the catchment scale. On the contrary, a drought is a creeping event that is evolving over months to seasons and can last for years. Droughts are occurring at regional, national, and even continental scales. Droughts are defined as a deficit of water with respect to normal conditions and are happening in all compartments of the hydrologic cycle such as the atmosphere (meteorological drought), rivers and groundwater (hydrological drought), and soils (agricultural drought).

Agricultural droughts can lead to severe socio-economic damages and humanitarian crisis (e.g., famine) because they have the potential to diminish crop yields. Impacts of these extreme events can be mitigated with the help of a seasonal prediction system that is able to forecast droughts several months in advance. This work is thus dedicated to the development of a seasonal soil moisture drought forecasting system over Europe. The meteorological forecasts by the North American Multi-Model Ensemble (NMME) are used within this prediction system to drive the mesoscale Hydrologic Model (mHM). The NMME product is selected because of the easy data access and the relatively large ensemble size. The prediction system developed in this work, however, provides enough flexibility to also incorporate other products such as TIGGE-LAM from the European Centre for Medium-Range Weather Forecasts (ECMWF).

The NMME dataset provides monthly precipitation and temperature forecasts. Daily values are also available but these require substantially higher computer resources, which are often not accessible to practitioners. Although the eight models comprised in this ensemble differ in their model structure (i.e., the selection and parameterization of implemented processes) they all represent the cycling of water and energy on Earth as a dynamic system. The mesoscale Hydrologic Model (mHM) is used in this study to represent the hydrologic cycle at the land surface. This choice is made because this spatially distributed model has been designed and calibrated to close the water balance at the outlet of a river catchment by reliably partitioning precipitation into evapotranspiration to the atmosphere and river discharge to the ocean. This partitioning is achieved by modeling soil moisture. The obtained soil moisture fields from mHM are then transformed to a quantile-based index to conduct a drought analysis.

The quality of a drought forecast depends on several factors. Major constituents of the seasonal prediction system are among others the meteorological forecasts obtained from the NMME. State-of-the-art drought prediction systems either evaluate the forecasts of single models or those obtained from the full ensemble mean (i.e., the average over all single model forecasts). Methods for increasing the skill of multi-model ensembles for reproducing extreme indices relevant for flood and drought modeling, which go beyond the evaluation of single models and the full ensemble mean, are investigated in this study. It is hypothesized in this work that a subset of models (subensemble) might give a better performance than both single models and the full ensemble mean. Different newly developed methods are investigated for selecting the best performing subensemble efficiently without requiring the evaluation of all possible subensembles.

Subensemble selection algorithms are tested for one of the latest collection of regional climate models made available by the ENSEMBLES project. The ability of these models to reproduce the spatio-temporal variability of eleven extreme precipitation and temperature indices, which are related to floods and droughts, is investigated. The estimated indices from these model simulations are compared against those obtained from the observations over Germany during the period from 1961 to 2000. The performance of single models and all possible subensembles is quantified using a rejection rate, which estimates the statistical significance of the obtained bias based on the Wilcoxon rank-sum test. The regional climate models, expectedly, exhibit smaller temperature bias than precipitation bias. The range of the bias for extreme precipitation indices is generally ±20%. Moreover, these regional climate models consistently overestimate or underestimate observations for a specific index. For example, long-term annual precipitation tends to be overestimated, whereas the maximum number of consecutive dry days is consistently underestimated.

For the ENSEMBLES regional climate models, the best possible subensemble showed smaller and less significant bias than all single models and the full ensemble mean in several parts of Germany. One of the proposed methods, the backward elimination algorithm, is able to find this subensemble efficiently without evaluating all possible subensembles. The identified subensemble contains only 6 out of the 13 considered regional climate models. Using this subensemble in favor of the full ensemble mean might lead to substantially different conclusions in climate change impact assessment studies.

The spatio-temporal resolution of a climate model is typically coarser than the resolution of a hydrologic model. For the drought prediction system developed here, the temporal resolution of the meteorological forecasts by the NMME is monthly values whereas the hydrologic modeling requires daily or hourly inputs. A newly developed stochastic method is employed to disaggregate the monthly precipitation fields into daily ones. This method is focusing on precipitation because of its skewed distribution function and its complex temporal and spatial covariance structure, which is induced by its intermittent occurrence. It is based on a multiplicative cascade approach that partitions one precipitation value at a low temporal resolution into two values at a higher resolution (e.g., one

bi-weekly value into two weekly ones) preserving the respective covariance structure at the corresponding temporal scale. This cascade procedure starts with monthly values and stops when daily values are obtained. The partitioning is achieved by drawing weights (i.e., multiplicative factors) from a distribution function that is estimated from the observations beforehand. These distribution functions are estimated for different levels of precipitation intensities and it is shown that they do not change significantly over time, which allows to apply this method during future periods, for example, in climate change impact assessment studies. It is hypothesized that the disaggregated daily fields obtained by employing this multiplicative cascade approach are statistically equivalent to the observed ones.

Although this method is focusing on the temporal disaggregation of precipitation, it also generates fields with a consistent spatial covariance structure. A newly developed sequential sampling technique termed “Anchor sampling” is used for this purpose. This method is applicable to grids of any resolution and extent. The “Anchor sampling” requires no assumption about the spatial isotropy of precipitation, which is a key property because orographic effects induce a non-isotropic spatial covariance structure. It can also be beneficial to any stochastic method that relies on the consistent spatial sampling of random fields (e.g., spatial downscaling schemes).

The stochastic disaggregation method is tested for a high-resolution precipitation dataset (i.e., 4×4 km^2 grid) over Germany ($\approx$ 357 000 km^2) during the period from 1950 to 2010. The evaluation of the method is conducted in a split-sample setup to evaluate the transferability to future periods using a calibration (1950-1990) and a validation (1991-2010) period. In general, distinctive location dependent distribution functions are reproduced with biases less than 5% during the evaluation period. Additionally, extreme precipitation indices are represented with biases less than 10%. These errors are less than those observed for the ENSEMBLES regional climate models (for the same indices) which highlights the added value of the proposed method. It is worth mentioning that the disaggregated precipitation fields have a temporal auto-covariance that compares well with the observed one. This is achieved without any explicit assumption about the auto-covariance structure of the disaggregated fields and is considered as an emergent property of the cascade model structure.

The seasonal drought prediction system then incorporates the backward elimination algorithm to identify the subensemble with the highest forecasting skill from the North American Multi-Model Ensemble (NMME). It also uses the newly developed stochastic temporal disaggregation method to downscale monthly precipitation forecasts to daily ones. Other forcing variables for the mesoscale Hydrologic Model (mHM) such as temperature are disaggregated by rescaling historic fields to have the same monthly values as the NMME forecasts. The forecasts obtained from the seasonal drought prediction system are contrasted with those of a simple statistical method based on the Ensemble Streamflow Prediction (ESP) approach. The ESP approach resamples past meteorological observations to create a forecast ensemble for driving mHM. This method thus provides a forecast that is only based on climatology and incorporates no knowledge about the actual dynamic development of the Earth system. The drought prediction system is evaluated during the period 1983-2009 with monthly forecasts starting at the beginning of each month for lead times up to 6 months. The spatial domain of the drought prediction system covers large parts of continental Europe. The forecasting skill is quantified employing the Equitable Threat Score (ETS), which combines the hit and false alarm rate of forecasted drought events. This metric requires reference soil moisture fields that

have been obtained using the freely available E-OBS dataset. It is hypothesized in this work that the NMME-based forecasts are exhibiting a higher ETS than the ESP-based ones.

The NMME-based soil moisture forecasts exhibit an ETS that is on average 69% higher than that of the ESP-based ones at a 6 month lead time. Results showed that there is a substantial spatial and temporal variability in drought forecasting skill ranging up to 40%. The drought forecasting skill is for both forecasting methods (i.e., NMME-based and ESP-based forecasts) higher in regions where reference soil moisture itself exhibits a high persistence. The perfect knowledge of the initial hydrologic conditions then leads to a high forecasting skill in these regions. Low drought forecasting skill is observed during periods of absence of drought indicating that the forecasts are overestimating drought extent during these periods. Drought forecasting skill is thus dependent on the state of drought development at the beginning of the forecast, which is contained in the initial hydrologic conditions.

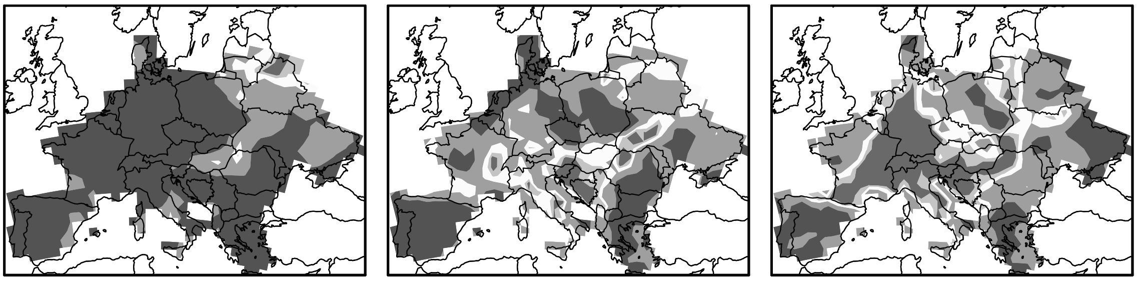
Among the NMME-based forecasts, the full ensemble mean exhibits a higher ETS than any identified subensemble and the best performing model. The skill of subensembles containing more than three models, however, is only 1% less than that of the full ensemble mean. But the number of realizations required in these subensembles is only 60% of that of the full ensemble mean. These subensembles thus could be useful for operational seasonal forecasting or practitioners that have limited computational resources. In general, the proposed drought prediction system that is based on the meteorological forecasts by the NMME should be taken into consideration in an operational system because it outperforms climatological forecasts based on the ESP approach over entire Europe at all lead times.

The methods presented in this work have not been tailored for the particular application of seasonal drought prediction. Instead, they are general and should be useful for other applications as well. For instance, the backward elimination method could be applied to other ensembles. Based on the results of this work, it is expected that this method could help to increase the utility of model ensembles in two ways: first, a higher performance than that of the full ensemble mean could be achieved by using subensembles; second, almost the same performance as that of the full ensemble mean could be achieved by subensembles, but at reduced computational costs.

The “Anchor Sampling” introduced in this work allows to generate spatial fields of normal distributed random numbers with a predefined spatial covariance structure. This method has only been employed for the temporal disaggregation of monthly precipitation to daily values. Follow-up studies should use the “Anchor Sampling” also for other applications such as, for instance, spatial downscaling of climate model outputs.

Chapter 1

Introduction



1.1 Background

Water is a vital element for all known life forms. There is around 1 385 000 000 km^3 of water on Earth (Chow et al., 1988) covering about 71% of the Earth's surface. The majority of this amount is stored in the oceans (96.5%), while only 2.5% is freshwater. A major portion of freshwater is comprised in Icecaps (68.5%). Only 30% of the global freshwater is contained in groundwater and even less (0.32% or 111 000 km^3) is stored in soils as soil moisture, rivers, lakes, and vegetation (Chow et al., 1988). It is this small portion of the total amount of water that is available to humans as a basis for their livelihood.

Water is constantly moving through different compartments of the Earth system such as the ocean, atmosphere, land surface, and subsurface. The largest fluxes within the global hydrologic cycle are the evaporation from the ocean to the atmosphere and the precipitation from the atmosphere to the Earth's surface. Roughly 20% of the global precipitation is over land (Bonan, 2008), which is the major input for driving river catchments.

Humans are changing the hydrologic cycle on land in various ways pursuing multiple objectives. These interventions are realized by building different hydrologic infrastructures. For instance, levees are constructed for protecting property from floods, dams are built to generate hydropower and to create reservoirs for drinking water supply, groundwater is pumped for freshwater supply, channels are built for transportation and irrigating agricultural land, and water is abstracted from rivers for industrial production such as cooling power plants. A profound knowledge of the hydrologic cycle at the catchment scale is required to appropriately design these infrastructures. Otherwise, mismanagement of water resources or failure of infrastructures is guaranteed. Hydrologic models are used to estimate states and fluxes of the hydrologic cycle at locations of interest within a catchment where no observations are available (Hrachowitz et al., 2013).

Important applications of hydrologic models are the prediction and hindcast of extreme events. The two most devastating hydrologic extremes are floods and droughts resulting in more deaths, yielding the highest economic losses, and affecting more people than any other natural hazard on a global scale (Table 2.2 in Sheffield and Wood, 2011). These two hydrometeorological phenomena are very different in their nature. A flood is a fast evolving event, which is spanning over days to weeks and is occurring at the catchment scale. It is often triggered by an extreme precipitation event. On the contrary, a drought is a creeping event, which is evolving over months to seasons and can last for years. Droughts are occurring at regional, national, and even continental scales. Droughts are defined as a deficit of water with respect to average conditions and are happening in all compartments of the hydrologic cycle such as the atmosphere (meteorological drought), rivers and groundwater (hydrological drought), and soils (agricultural drought) as discussed in Sheffield and Wood (2011).

Agricultural droughts have devastating impacts on human societies because they have the potential to diminish crop yields among other negative environmental effects (e.g., the reduction of the carbon sink strength of forest ecosystems as shown by Piayda et al., 2014). For example, the European 2003 drought caused alone in Germany socio-economic damages amounting to more than EUR 1.5 bn (COPA-COGECA), whereas the 2010/11 drought in the Horn of Africa led to a severe humanitarian crisis affecting more than 12 million people (Relief, 2011). A reconstruction of the 2003 European drought development is exemplary shown in Figure 1.1. Areas exhibiting a drought are marked as yellow and orange regions in this Figure. Almost no grid cell has been under drought during February 2003. The drought then first started to develop in Eastern Europe as well as Italy during May 2003. The peak extent is observed during August 2003 where major parts of Central Europe are under severe drought (i.e., are exhibiting a low soil moisture value that has less than 5% probability of occurrence).

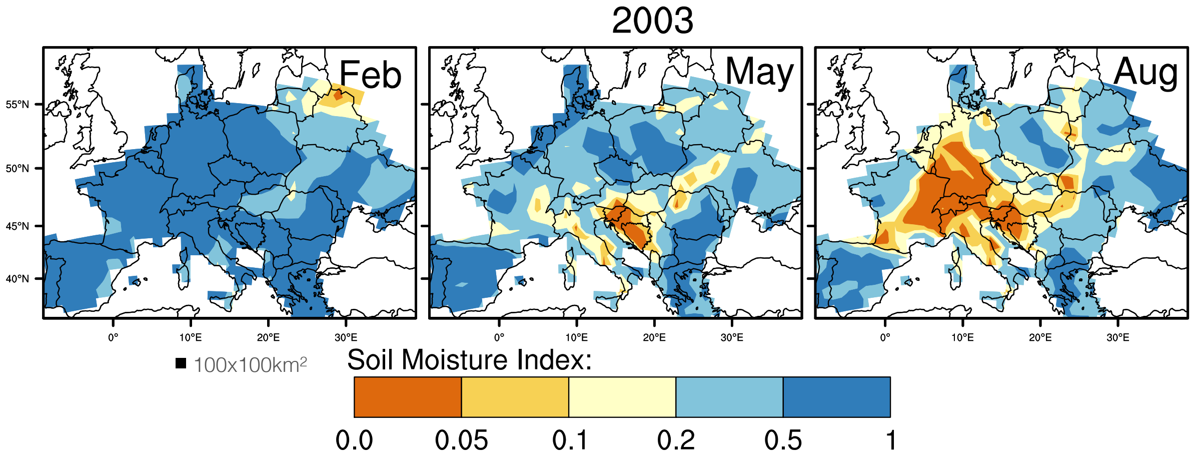


Figure 1.1: The development of the 2003 European drought is illustrated by depicting the drought extent for February, May, and August of 2003. A drought is occurring when the soil moisture index is below 0.2 indicating a low fractional soil moisture value that is observed less than 20% of the time. These fields are obtained by driving the mesoscale Hydrologic Model (mHM) with the observation-based E-OBS dataset.

Agricultural droughts are developing at seasonal time scales as illustrated in Figure 1.1. Hence, a seasonal drought prediction system that forecasts these events several months in advance could help to adapt to these extreme events and mitigate their impact. This work is thus aiming at the development of a seasonal drought prediction system for Europe.

1.2 Seasonal Drought Prediction System

The general framework of a seasonal drought prediction system is depicted in Figure 1.2. This framework is consisting of three major components. The first component is devoted to the meteorological forcing that is required to drive the hydrologic model, most importantly daily temperature and precipitation. The second component is the hydrologic model that transforms the meteorological input into hydrologic states and fluxes at the land surface. In this work, the mesoscale Hydrologic Model (mHM, Samaniego et al., 2010; Kumar et al., 2013a) is used. The variable of interest in this work is the simulated soil moisture. The third component is the post-processing of simulated soil moisture to a quantile-based soil moisture index for the drought analysis (bottom box in Figure 1.2). The structure of this framework is very similar to those used in other studies such as Schaake et al. (2007b), Luo and Wood (2007), and Yuan et al. (2015).

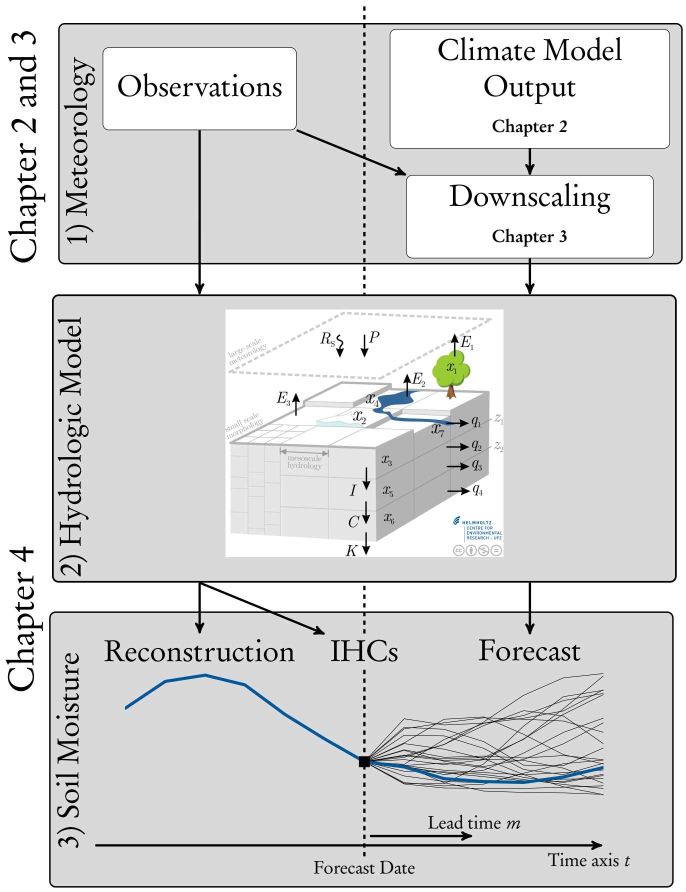


Figure 1.2: Schematic representation of a drought prediction system. This Figure is divided vertically into an observation-based part (left of vertical dashed line) and a forecasting part (right of vertical dashed line). The observation-based part is comprised of: first, an observation-based meteorological dataset; second, a hydrologic model that is driven with these observations. Third, the simulated soil moisture is used for reconstruction during historic periods and to provide initial hydrologic conditions (IHCs) for the forecasts. The forecasting part (i.e., the right hand side of this figure) consists of: first, a meteorological forecast obtained from a climate model (or numerical weather prediction model); second, a downscaling of climate model outputs that requires also the meteorological observations; third, a hydrologic model of the land surface that is driven with the downscaled meteorological forecasts to provide a soil moisture forecast. This work is mainly focusing on the evaluation of climate model outputs with respect to extreme meteorological indices (Chapter 2) and the temporal downscaling of meteorological variables (i.e., disaggregation method introduced in Chapter 3). The evaluation of the drought prediction system is then presented in Chapter 4.

There are two datasets of meteorological forcings required to setup the drought prediction system. The first is an observation-based dataset for past periods and the second is the actual meteorological forecast (top box in Figure 1.2). The observation-based dataset is required mainly to spin-up the hydrologic model and to create the initial hydrologic conditions (IHCs). If the forecast period is in the past, then the observation-based soil moisture can be also used as a reference to assess the skill of the soil moisture predictions. These soil moisture predictions are obtained from the meteorological forecasts, which are typically derived from climate models or numerical weather prediction models that represent the Earth's energy and water cycle as a dynamic system. In this work, the seasonal meteorological forecasts are obtained from the North American Multi-Model Ensemble (NMME) that comprises realizations of eight global climate models. The NMME product is selected because of the easy data access and the relatively large ensemble size. Any other product such as TIGGE-LAM from the European Centre for Medium-Range Weather Forecasts (ECMWF; Buizza, 2015) could, however, also be used.

The spatio-temporal resolution of the meteorological forecasts obtained from climate model simulations is typically coarser than the resolution of the hydrologic modeling. For example, the NMME product provides monthly meteorological forecasts whereas mHM requires daily or hourly inputs. Due to this fact, climate model datasets need to be downscaled, which is a kind of pre-processing step included in the first component of the drought prediction system (top box in Figure 1.2).

The second component of this drought prediction system consists of the mesoscale Hydrologic Model (mHM, Samaniego et al., 2010; Kumar et al., 2013a), which is driven with the downscaled daily forcing data to estimate gridded soil moisture fields. mHM is a spatially explicit distributed hydrologic model in which hydrologic processes are conceptualized similarly to these of other existing large-scale models like VIC-3L (Liang et al., 1996) and WaterGap (Döll et al., 2003). The included processes are canopy interception, snow accumulation and melt, soil moisture and infiltration, runoff generation and evapotranspiration, deep percolation and base flow, and flood routing between grid cells. These processes are parameterized in a unique way employing the multiscale parameter regionalization (MPR) technique that allows for parameter transferability across spatial resolutions and locations. Further details on the model and the parameter regionalization are given in Samaniego et al. (2010) and Kumar et al. (2013a).

It is worth mentioning that the climate models also include a land surface model, which provides the lower boundary fluxes to the atmosphere over land. These models conceptualize hydrologic processes such as soil infiltration, percolation, and snow pack in a similar way as mHM. The fundamental difference between these two kind of models is that mHM has been calibrated to match the observed river discharge at the outlet of a river catchment which is not the case for land surface models in general. It is thus hypothesized that mHM correctly represents the partitioning of precipitation into evapotranspiration and river discharge at the catchment scale. This partitioning is achieved by modeling soil moisture and it is assumed that the simulated soil moisture dynamics obtained by mHM are a valid representation of the true ones.

The third component is the post-processing of soil moisture. Previous studies have shown that soil moisture dynamics strongly depend on the structure of the model employed within the simulation (Koster et al., 2009; Wang et al., 2011). This makes it very difficult to compare soil moisture among different models and also to evaluate it within a drought study. Simulated soil moisture is thus often transformed to a quantile-based soil moisture index, which quantifies the probability of occurrence at a given point in time (a.o., Sheffield et al., 2004; Vidal et al., 2010; Samaniego et al., 2013). The

20% quantile (i.e., the soil moisture value that is undercut 20% of the time) is then used as a drought threshold. In this work, the approach outlined by Samaniego et al. (2013) is adopted.

The quality of drought forecasts depends on several factors. The uncertainty of the different components of the prediction system (e.g., the meteorological forcing, the hydrologic model structure, and the kind of index chosen in the post-processing) will contribute differently to the uncertainty of drought forecasts. The evaluation in this work is focusing on the input uncertainty of the drought prediction system because this might be responsible for the largest uncertainty within the soil moisture forecasts. The uncertainty in other components is, however, not negligible. For example, the uncertainty in the initial hydrologic conditions (IHCs) also contributes substantially to the uncertainty of soil moisture forecasts (Wood and Lettenmaier, 2008). In this work, perfect knowledge of IHCs, which are taken from a reference dataset, is assumed. The investigation of the validity of this assumption is subject to follow-up studies. With respect to the input uncertainty, there are mainly two contributors within a drought prediction system. Both of these are investigated in this work. The first is the quality of the meteorological forcing made available by a multi-model ensemble of climate models. The second is the quality of the fields obtained from the downscaling procedure. The next two sections present the current state-of-the-art techniques how these two issues are addressed and which improvements are required for a seasonal drought prediction system for Europe.

1.2.1 Evaluation of Multi-Model Ensembles of Climate Models

This work employs the forecasts made available by the North American Multi-Model Ensemble (NMME, Kirtman et al., 2014). This dataset provides 101 monthly precipitation and temperature forecasts from different global climate models. There is substantial uncertainty associated with this ensemble dataset, which is a common feature of climate model simulations. These uncertainties occur mainly because of two reasons. First, climate models are of different complexity representing different processes (Flato et al., 2013). Second, there is internal climate variability that introduces some uncertainty even if the same model is evaluated over the same domain with small perturbations in the initial atmospheric conditions (Deser et al., 2012). Both of these characteristics cannot be modified by a practitioner to increase the skill of climate model outputs for a particular hydrologic application of interest. Instead, a practitioner can only increase the skill by post-processing this data.

The uncertainty in the meteorological forcing then translates to an uncertainty in the soil moisture forecasts which is exemplary shown in Figure 1.3 for NMME-based forecasts. There is a considerable spread among the forecasts obtained from the different models. Previous studies have evaluated the skill of NMME-based hydrologic predictions by evaluating either the skill of single models or the full ensemble mean (Yuan and Wood, 2013; Mo and Lettenmaier, 2014; Yuan et al., 2015, a.o.). This is a standard approach in the evaluation of multi-model ensembles regardless of the dataset and application of interest (Giorgi and Mearns, 2002; Weigel et al., 2010; Soares et al., 2012; Schindler et al., 2012). It can, however, be observed in Figure 1.3 that there are some models that show a higher agreement with respect to the reference soil moisture anomaly than others (e.g., CMC1 and COLA). This asks for methods that try to provide a higher skill than that of the full ensemble mean.

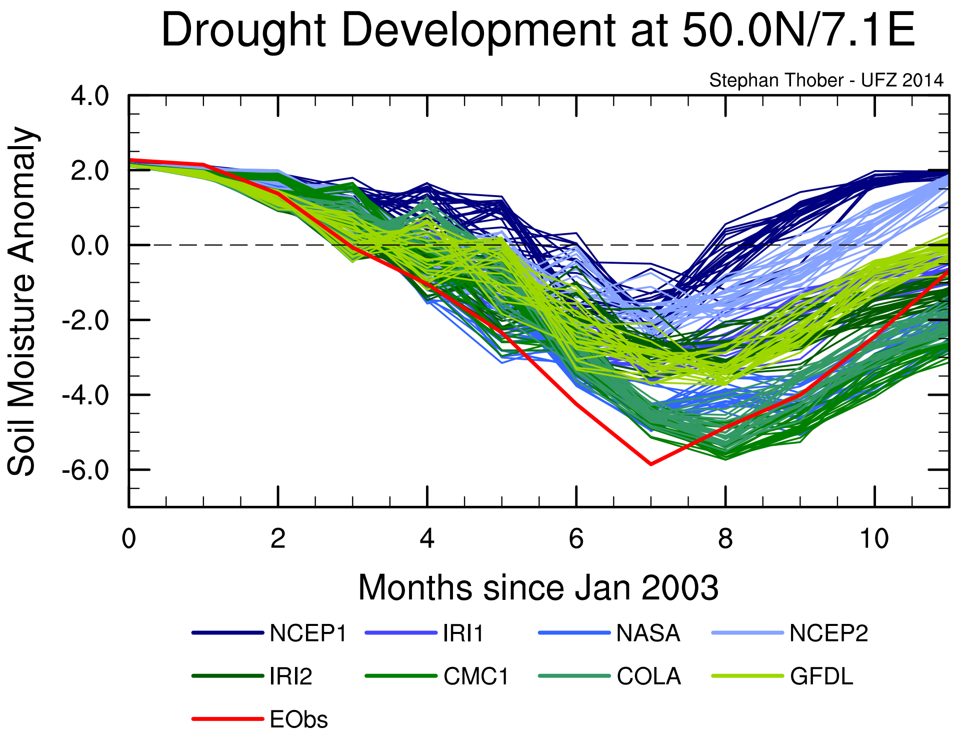


Figure 1.3: Standardized soil moisture anomalies for the period from January to December 2003. The red line depicts the reference anomaly obtained by the E-OBS dataset. The blue and green lines show the soil moisture anomalies of NMME-based forecasts with one month lead time.

Among the different methods that aim at improving the ensemble skill, weighted ensemble averaging techniques have become a well accepted practice in climate change impact assessment studies (Giorgi and Mearns, 2002; Doblas Reyes et al., 2005; Weigel et al., 2010). These averaging schemes assign higher weights to models exhibiting a higher performance with respect to a selected metric as compared to the performance of other models. The weighted average should then provide a higher performance than a simple arithmetic multi-model average. This method is, however, not applicable to a variable like precipitation, which is a fundamental input variable for the drought prediction system. This is due to the fact that the weighted average of this variable lacks key characteristics of the original fields like, for example, a skewed distribution function and a strong spatio-temporal intermittency.

An alternative approach would be to only select a subset of models (subensemble), which might exhibit a higher performance as compared to the full ensemble mean. For example, only select the COLA and CMC1 model in Figure 1.3. Such an approach has not been investigated yet which might be due to the large number of potential subensembles to evaluate (this number increases proportional to the factorial of the number of given models in the ensemble). This work aims at investigating the performance of subensembles as well as at proposing an efficient method to find the best performing subensemble of a given multi-model ensemble.

The use of subensembles in favor of the full ensemble mean could be beneficial in multiple ways. First, a subensemble could exhibit the highest possible performance, which would increase the performance as compared to that of single models and the full ensemble mean. This would have direct implications for climate change impact assessments because subensembles could, for instance, provide a less biased estimate. Second, subensembles could provide the same performance as the full ensemble mean, but at a reduced number of considered model realizations. Practitioners that have limited computational resources could exploit this and use the subensemble as a surrogate for the full ensemble mean.

Multi-model ensembles of climate models are nowadays routinely used for the evaluation of the impact of climate change on extremes like floods and droughts. Subensemble selection methods should be general, such that they can be useful for this wide range of applications. The performance of subensembles and the methods for efficiently selecting them is thus investigated for another dataset than the North American Multi-Model Ensemble (Chapter 2). This dataset consists of thirteen regional climate models which are obtained from the ENSEMBLES project (van der Linden and Mitchell, 2009). These models are tested for their ability to reproduce several extreme temperature and precipitation indices that are relevant for floods and droughts. Different subensemble selection methods are evaluated for this dataset and the most efficient one is then applied within the drought prediction system.

1.2.2 Downscaling Techniques for Meteorological Variables

The spatial and temporal resolution of climate model outputs is typically too coarse for hydrologic assessment studies. For example, the spatial resolution of state-of-the-art global climate models is typically in the order of 0.5° to 4° (Taylor K. et al., 2012), which does not provide any spatial variability of meteorological variables (e.g., precipitation and temperature) at the catchment scale needed for spatially distributed hydrologic modeling. Additionally, the substantial storage requirements often allow providing the meteorological climate model outputs only at a low temporal resolution of monthly values although the internal time steps are in the order of seconds to minutes. As a consequence, low-resolution climate model outputs have to be disaggregated in time and downscaled in space to drive a hydrologic model at a high-resolution, which is schematically shown in Figure 1.4.

Techniques that increase the spatial and temporal resolution of climate model outputs are therefore required (von Storch, 1999; Maraun, 2013; Thober et al., 2014). Notably, a method that increases the temporal resolution of a dataset has to ensure a consistent spatial structure at the higher temporal scale as well. Vice versa, a spatial downscaling also has to preserve a consistent temporal variability. The reason is that the spatio-temporal structure of meteorological variables is highly dependent on both the considered spatial and temporal resolution. For example, precipitation is less intermittent in space at a low temporal resolution (e.g., monthly values) than at a high one (e.g., daily values). This also implies that the complexity of spatial downscaling depends on the temporal resolution at which it is applied and, in general, is simpler at a lower one than at a higher one (Figure 1.4). On the contrary, the complexity of temporal disaggregation does not vary with spatial resolution because it has to add spatial intermittency in any case. In the past decades, a variety of downscaling techniques have been investigated and extensive reviews can be found in (Wilks and Wilby, 1999; van der Linden

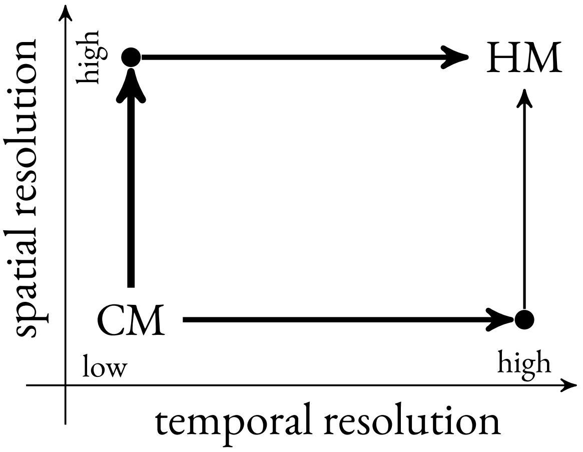


Figure 1.4: Outputs of climate models (CM) are typically available at lower spatial and temporal resolutions than the input required for hydrologic models (HM). Temporal disaggregation schemes such as those presented in Chapter 3 increase the temporal resolution of CM to that of HM (horizontal arrows). Spatial downscaling such as those discussed in Section 5.2 increase the spatial resolution (vertical arrows). The thickness of the arrows corresponds to how simple a particular step is in comparison to other ones.

and Mitchell, 2009; Maraun et al., 2010, a.o.). Downscaling techniques can be categorized into two major classes: dynamical and statistical downscaling.

Dynamical downscaling is carried out using regional climate models, which use essentially similar representations of physics to describe the state of the atmosphere and land surface as global climate models. Such techniques are not considered in this work because their computational demand is comparable to that of global climate models. This high computational demand can only be met by super-computer facilities, which are not available to hydrologic practitioners and modelers in general.

A computationally inexpensive alternative to dynamical downscaling is statistical downscaling. These downscaling techniques exploit statistical relationships between low-resolution predictors and high-resolution predictands. Climate model outputs are often taken as low-resolution predictors. An estimate for the high-resolution predictor is also required to establish the statistical function employed in the downscaling. This estimate is often obtained by using observation-based datasets of meteorological variables. For this reason, observation-based datasets contribute to the downscaling of climate model data as shown in Figure 1.2 (top box). A plentitude of statistical relationships has been used in previous research such as quantile mapping (Wood et al., 2002), copulas (van den Berg et al., 2011), and multiplicative cascade models (Güntner et al., 2001). A comprehensive review can be found in Maraun et al. (2010). In this work, a multiplicative cascade approach that has been used for the temporal disaggregation of precipitation at sub-daily scales (Güntner et al., 2001) is adapted for the disaggregation of monthly to daily precipitation.

Weather generators are algorithms providing sequences of random numbers that have the same statistical properties (i.e., statistical moments) as observations and constitute one subclass of statistical downscaling methods (Maraun et al., 2010). These techniques generate an ensemble of feasible high-resolution variables given a low-resolution predictor, which quantifies the uncertainty within the high-resolution predictand realistically. Among the different variables considered in weather generators like temperature, wind speed, and humidity, the correct representation of statistical moments is particularly challenging for precipitation. This is due to the fact that precipitation exhibits a highly skewed distribution function for positive values and, additionally, a high spatial intermittency. The correct representation of this variability and spatial structure is fundamental for a drought prediction system. For instance, if the disaggregated time series would distribute precipitation too equally in time, then there would be little chance to accurately model high deviations from normal conditions such as droughts (and also floods).

These challenges are addressed in this work by investigating a disaggregation method that increases the temporal resolution of a precipitation dataset and simultaneously generates realistic spatial fields on grids of any resolution and extent (Chapter 3). A newly proposed sequential sampling method is developed in this work for this purpose. This sampling method does not require the assumption of isotropic fields, which is often not fulfilled due to orographic effects. The disaggregation of other variables (e.g., temperature) uses an approach that rescales historic fields to match the forecasted one. This approach preserves the historic spatio-temporal pattern of observed fields. Given the fact that the North American Multi-Model Ensemble (NMME) provides monthly estimates of precipitation and temperature, the investigated downscaling method will focus mainly on increasing the temporal resolution of this dataset and no spatial downscaling is applied in this work. The drought prediction system is evaluated at a continental scale for which the relatively low spatial resolution of the NMME outputs ($1^\circ \times 1^\circ$) still allows to perform a spatially distributed analysis. The discussion in Section 5.2, however, also examines potential links of the presented disaggregation method to spatial downscaling.

1.3 Objectives and Research Questions

The main objective of this work is to develop and to benchmark a seasonal drought prediction system for Europe. The preceding section has presented the main components of such a framework, which are also shown in Figure 1.2. The focus of this work is in particular the processing of the meteorological forcing. This work aims at enhancing state-of-the-art downscaling techniques and methods for evaluating multi-model ensembles.

One approach to increase the skill of the prediction system is to investigate methods that evaluate the ability of a given multi-model ensemble more comprehensively than relying on the performance of either single models or the full ensemble mean. This comprehensive evaluation is achieved by analyzing the performance of subensembles. Methods for selecting appropriate subensembles should be general given the extensive applications of multi-model ensembles in current research. The development of subensemble selection methods is thus carried out for another multi-model ensemble of climate models than the one employed within the drought prediction system. The analysis is focusing on meteorological extremes because these will translate to hydrologic extremes (e.g., floods and droughts) while propagating through the land surface compartment of the hydrologic cycle. The main research questions related to this topic are as follows:

1.1) *Are current climate models able to reproduce the spatio-temporal variability of observed extreme statistics of precipitation and temperature?*

1.2) *Is a subensemble of climate models more suitable to reproduce meteorological extremes than the full ensemble mean of all considered models?*

Chapter 2 is devoted to give an answer to these research questions. This chapter is published as Thober and Samaniego (2014). The best performing method for efficiently selecting subensembles will then be also implemented in the seasonal drought prediction system.

The monthly estimates obtained from the North American Multi-Model Ensemble have to be disaggregated to daily values for the hydrologic modeling. There are two main requirements for the disaggregation technique that are imposed by the drought prediction system for Europe. First, the downscaled precipitation fields should reproduce extreme dry spells reasonably well. Second, the proposed method should be applicable to large grids and preserve the realistic spatio-temporal structure of precipitation as much as possible. In this work, a multiplicative cascade approach (Güntner et al., 2001) is adapted for this purpose and jointly used with a newly developed sequential sampling method, termed "Anchor Sampling". The main research questions are as follows:

2.1) *Is a multiplicative cascade approach able to disaggregate monthly precipitation estimates to daily values preserving observed statistical properties, in particular wet and dry extremes?*

2.2) *Is the "Anchor Sampling" able to generate precipitation fields with a realistic spatial covariance structure?*

Chapter 3 presents an analysis of the newly developed disaggregation method over Germany to address these research questions. This chapter is published as Thober et al. (2014).

The subensemble selection algorithm introduced in Chapter 2 and the temporal disaggregation method presented in Chapter 3 are then used within the seasonal drought prediction system for Europe. The main research questions with respect to the development and the benchmark of this prediction system are as follows:

3.1) *Are drought forecasts based on the North American Multi-Model Ensemble more skillful than forecasts based on a simple statistical forecasting method (i.e., the Ensemble Streamflow Prediction approach) over larger parts of the European domain?*

3.2) *Is drought forecasting skill distributed uniformly in space and time?*

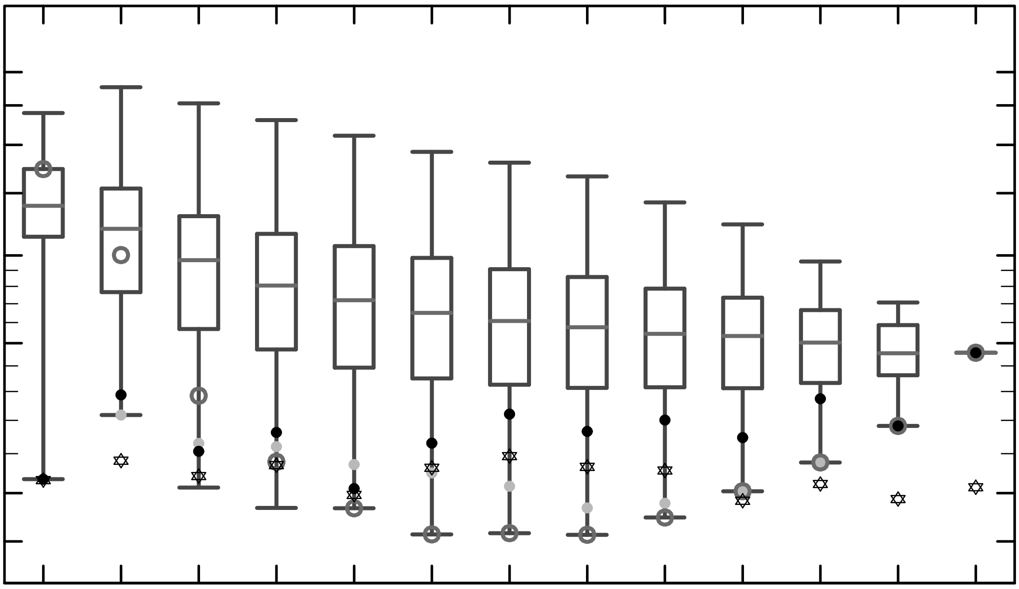
Chapter 4 presents the detailed description of the drought prediction system and the analysis of the benchmark. This chapter is submitted as Thober et al. (2015).

Subsequent to the presentation of these three single studies, Chapter 5 summarizes the main results and discusses links between these that have not been addressed within the single studies. This chapter also contains a discussion on limitations of the presented methods that should be considered in follow-up studies.

the three papers

Chapter 2

Robust Ensemble Selection by Multivariate Evaluation of Extreme Precipitation and Temperature Characteristics



Aknowledgments:
This study was carried out within the context of the WESS and REKLIM projects and was kindly supported by the Helmholtz Impulse and Networking Fund through Helmholtz Interdisciplinary School for Environmental Research (HIGRADE). The authors also acknowledge the collaboration of the German Weather Service DWD for providing the observational data set employed in this study. The ENSEMBLES data used in this work was funded by the EU FP6 Integrated Project ENSEMBLES (Contract number 505539) the support of which is gratefully acknowledged. We additionally acknowldege DMI for maintaining the RCM data portal. Furthermore, the authors would like to thank Rohini Kumar and Juliane Mai for vivid discussions and careful proof reading. The authors would also like to thank the Editor and three anonymous reviewers for their valuable suggestions and critique on the manuscript.

published as:
Stephan Thober and Luis Samaniego, "Robust Ensemble Selection by Multivariate Evaluation of Extreme Precipitation and Temperature Characteristics", Journal of Geophysical Research: Atmospheres, 119(2):594–613, 2014, ISSN 2169-8996, doi:10.1002/2013JD020505

2.1 Abstract

Extreme hydro-meteorological events often cause severe socio-economic damage. For water resources assessments and policy recommendations, future extreme hydro-meteorological events must be correctly estimated. For this purpose, projections from Regional Climate Models (RCMs) are increasingly used to provide estimates of meteorological variables such as temperature and precipitation. The main objective of this study is to investigate whether a full ensemble or a subset of RCMs reproduces the spatio-temporal variability of observed extremes better than single models. The implications for policy recommendations and impact assessments are then discussed. In particular, the key conditions under which a subset of RCMs could be used for impact assessments are examined. Temperature and precipitation fields of 13 ENSEMBLES RCMs are compared against observations from Germany between 1961 and 2000. Eleven indices characterizing extreme meteorological events were selected for this comparison. The ability of the individual RCMs is estimated based on an overall score and a rejection rate. The former quantifies the biases of these indices. The latter estimates the mean statistical significance quantified by the Wilcoxon rank-sum test. The performance of all possible combinations of RCMs is investigated. Computationally feasible algorithms for finding the best-performing subensemble are also presented and evaluated. One of the proposed algorithms is able to find subensembles with the lowest rejection rate, which are useful for either policy recommendations or impact assessments. These subsets of RCMs showed smaller and less significant bias than single RCMs or the full ensemble over several regions.

2.2 Introduction

Extreme hydro-meteorological events are the major cause of major natural disasters according to the United Nations. According to the insurance group Munich RE, the overall losses due to flood- and drought-related events amounted to USD 570 bn globally for the year 2008. In Germany, extreme weather conditions are causing severe socio-economic damage. The flood along the Elbe river in 2002, for example, caused economic damage totaling EUR 9.4 bn. During the summer heat wave of 2003, there were approximately 7,000 more deaths than during an average summer (Schuchardt et al., 2008). Globally, the discharge regimes in most river basins and their top-soil moisture patterns are expected to be altered by climate change. Specifically in Germany, the flood risk in winter and spring is expected to rise, low waters in summer will become more frequent, and groundwater tables will change, with possible consequences for the supply of drinking water (Schuchardt et al., 2008).

Water resource management that is cost effective and risk averse requires, among other things, estimates of the precipitation, runoff, soil moisture, and groundwater recharge at given locations as well as an estimated probability of extreme events affecting variables such as precipitation or runoff, among others.

Continuous simulation models (Samaniego et al., 2010), extreme value and extreme excess theories (Davison and Smith, 1990; Koutsoyiannis et al., 1998) have been used in modern hydrologic science and engineering to address these challenges. There is strong evidence, however, that climate change has altered hydro-climatic regimes all over the world, changing precipitation and temperature. As a result, the estimates of the probabilities of the occurrence of extreme events are very likely underestimating the future magnitude of design events for a given level of risk.

Decision makers, on the other hand, urgently require insights on how to conceive adaptation and mitigation strategies for the regions that are expected to undergo a significant hydro-climatic change in the near future. To fulfill these expectations, a number of studies have been carried out in recent years (Lopez et al., 2009; Pelt and Swart, 2011; Bormann et al., 2012; Matrosov et al., 2013). The common assumption of these studies is that the current generation of Regional Climate Models (RCM) is able, at least in part, to resolve the observed dynamic of the meteorological forcings required to drive hydrologic models that aim at generating predictions for variables of interest (e.g., streamflow discharge). Many authors have noted, however, that the probability density functions (pdfs) of precipitation and air temperature obtained from RCMs do not quite match the observed distributions. As a result, the so-called "bias correction" method has been introduced to compensate for RCM deficiencies (Hay et al., 2000; Li et al., 2010; Piani et al., 2010; Lafon et al., 2012). A number of procedures have been introduced to downscale RCM data to the desired spatial resolution to generate forcing data at the scale of the impact models. Comprehensive reviews of these methods can be found in the literature (Maraun et al., 2010).

If RCM "weather" is to be used for policy recommendations or as forcing for impact assessments, then it should fulfill a number of conditions. Among the most important ones is the condition that the synthetic weather generated by an RCM should exhibit good estimates for the extreme characteristics of precipitation and air temperature that critically impact hydrology. The importance of this condition is that if these RCM outputs do not fulfill this criterion, then any downscaled product or impact assessment product (e.g., discharge) derived from them will also fail to align with the observations.

Hence, the evaluation of RCM outputs is of increasing interest. Extensive research has been undertaken to investigate the reliability of RCMs in modeling the current climate (period from 1961 to 1990). For example, there are a number of studies that have focused on the evaluation of large-scale circulation patterns, probability density functions, and internal, seasonal, and inter-annual variability (Jaeger et al., 2008; Sanchez-Gomez et al., 2008; Rauscher et al., 2010; Boberg et al., 2010). In general, studies aiming to evaluate the ability of RCMs to reproduce or project the spatio-temporal patterns of extreme characteristics (Frei et al., 2006; Herrera et al., 2010; Schindler et al., 2012; Soares et al., 2012) and to reproduce key hydrologic variables (e.g., streamflow) (Hay et al., 2002; Music and Caya, 2007) only have relied on quantifying the magnitude of the bias of the considered statistics. The level of significance at which an RCM is able to reproduce extreme characteristics of precipitation and air temperature (here after referred to as indices) was mostly evaluated by parametric extreme value distribution theory (Beniston et al., 2007; Fowler et al., 2010). To our knowledge, however, non-parametric significance tests have not been employed for this purpose, although they do not require assumptions related with the sampling distribution of the test statistic, which is highly advantageous for analyzing extreme indices.

Because it is very unlikely that a single RCM could model all indices equally well, using ensemble averaging techniques to generate the desired outputs instead of a single model output has become a well-accepted practice in the scientific community (Giorgi and Mearns, 2002; Doblas Reyes et al., 2005; Weigel et al., 2010). Ensemble averaging using equal weighting is the standard approach for investigating the performance of a multi-model ensemble (MME). Other possibilities are weighted averaging schemes such as the "Reliability Ensemble Averaging" (REA) approach (Giorgi and Mearns, 2002), which estimates weights for each model based on its "performance" during past periods and "convergence" during future ones. Statistical methods such as hierarchical ANOVA models (Sansom et al., 2013) and Bayesian frameworks (Tebaldi et al., 2005; Smith et al., 2009; Chandler, 2013) have also been investigated. Evaluating the ability of RCMs to reproduce multiple extreme indices with a Bayesian framework would imply fitting statistical models to every grid cell and index. As a consequence, such an approach would be computationally challenging for regional assessments.

Current studies, in general, are only investigating the performance of either single RCMs or the full ensemble average (Giorgi and Mearns, 2002; Weigel et al., 2010; Soares et al., 2012; Schindler et al., 2012). The performance of subensemble averages is often not considered. Therefore, it is not known whether a subensemble average could outperform the full ensemble average. Rigorously tested methods to select subensembles have not been investigated too. Subensembles could be highly beneficial for impact assessments because they could provide a good selection of single RCMs that can be used as forcing for hydrologic impact models. Subensembles could also be useful for policy recommendations because they could provide a less biased multi-model average of indices than that of the full ensemble.

Within this framework, this study aims to address three fundamental research questions. First, are current RCMs able to reproduce the spatio-temporal variability of observed extreme statistics of precipitation and temperature in Germany? If the null hypothesis associated with the first research question would be rejected at a significance level of 5%, for example, it would be extremely useful to know whether these simulations could be used for reliable regional hydrologic projections, and if so, where. Second, is a model ensemble better able to reproduce extreme statistics than single members, as is commonly hypothesized (Jun et al., 2008; Weigel et al., 2010; Suh et al., 2012)? Third, is a subensemble average of RCMs more suitable to reproduce extreme statistics than the full ensemble average of all RCMs? If so, is there a simple procedure to find the most robust subset of RCMs?

In this study, the performance of RCMs with respect to multiple extreme indices (thus multivariate) is investigated using an overall score and a rejection rate. The former quantifies the biases of these indices. The latter estimates the mean statistical significance quantified by the non-parametric Wilcoxon rank-sum test.

Moreover, simple non-weighted and weighted averaging schemes (Giorgi and Mearns, 2002) are applied to generate a multi-model average of a given extreme index, not only for the full ensemble but also for all possible combinations of models. Exploring the whole space of combinations of models might help to extract and synthesize the relevant information from RCMs (Knutti et al., 2010). Computationally feasible algorithms for finding the best-performing subensemble are also presented and evaluated.

Considering that the aim of the present study is to investigate whether subensemble averages are able to reproduce extreme indices better than the full ensemble average, the original RCM outputs should be used without any numerical transformation. For this reason, bias-correction techniques are not considered in this study.

Furthermore, the availability of highly resolved observational data sets is crucial for the evaluation of RCMs, as demonstrated by previous research (Herrera et al., 2010; Soares et al., 2012; Rauscher et al., 2010). In the present study, it is also investigated whether high-resolution precipitation and temperature data sets provided by the German Weather Service (DWD) are more suitable for the evaluation as compared to coarse-resolution products such as EOBS (Haylock et al., 2008).

2.3 Data Sets and Study Domain

This study was carried out over a spatial domain ranging from 10°E to 15°E and 50°N to 56°N, which covers the entire territory of Germany. The terrain is characterized by lowlands in the north, small mountain ranges in the central Germany ($\leq$ 1,000 m altitude), and the Alps at the southern border of Germany ($\leq$ 3,000 m altitude). The climate is mild with a mean annual temperature of approximately 8.5° C. The long-term annual precipitation ranges from 700 mm in the lowlands up to 2,000 mm in the Alpine regions. The numerical analysis of this study was conducted for the period from 1961 to 2000.

The RCM simulations carried out in the ENSEMBLES project were selected because they are the most recent and comprehensive collection of data sets over the whole of Germany for the period from 1961 to 2000 (van der Linden and Mitchell, 2009). 13 RCM data sets were obtained from the ENSEMBLES project, as displayed in Table 2.1. These RCMs were forced by the ERA 40 reanalysis data (Uppala et al., 2005) and are available at a 25 km spatial resolution. However, because the spatial projections are not the same, all model outputs were remapped to the observational grid (transverse Mercator projection). Conservative remapping (Jones, 1999) as implemented in the Climate Data Operators (Schulzweida, 2013) was employed for this purpose. This remapping procedure conserves areal fluxes (e.g., precipitation), which is a crucial requirement for the analysis of the RCMs.

Symbol	Acronym	Institute	Model	Reference
A	C4IRCA3	Community Climate Change Consortium for Ireland	RCA3	Kjellström et al. (2005)
B	CHMIALADIN	Czech Hydro-Meteorological Institute	ALADIN	Farda et al. (2010)
C	CNRM-RM4.5	Centre National de Recherches Meterologiques	RM4.5	Radu et al. (2008)
D	DMI-HIRHAM5	Danish Meteorological Institute	HIRHAM	Christensen et al. (2006)
E	ETHZ-CLM	Swiss Institute of Technology	CLM	Jaeger et al. (2008)
F	ICTP-REGCM3	Abdus Salam International Centre for Theoretical Physics	REGCM3	Pal et al. (2007)
G	KNMI-RACMO2	Koninklijk Nederlands Meteorologisch Instituut	RACMO	van Meijgaard et al. (2008)
H	METNOHIRHAM	The Norwegian Meteorological Institute	HIRHAM	Haugen and Haakenstad (2005)
I	METO-HC HadRM3Q0	Hadley Center / UK Met Office	HadRM3 Q0	Collins et al. (2006)
J	MPI-M-REMO	Max Planck Institut fuer Meteorologie	REMO	Jacob et al. (2001)
K	OURANOSMRCC4.2.3	Consortium on Regional Climatology and Adaptation to Climate Change (Montreal)	CRCM	Music and Caya (2007)
L	RPN-GEMLAM	Environment Canada	GEMLAM	Zadra et al. (2008)
M	SMHIRCA	Swedish Meteorogical and Hydrological Institute	RCA	Samuelsson et al. (2011)

Table 2.1: Regional Climate Models incorporated in this study from the ENSEMBLES project (van der Linden and Mitchell, 2009)

Two data sets were utilized for the observed precipitation, namely, the 25 km European product EOBS (Haylock et al., 2008) and the 1 km gridded precipitation product REGNIE from the German Weather Service (www.dwd.de), which takes into account more than 5,000 meteorological stations across Germany. The latter was aggregated to 25 km spatial resolution to make it comparable with both the available RCM data and the EOBS data set. As for precipitation, two data sets for the observed temperature were used in this study. Those were from EOBS (25 km) and data from 1,160 German Weather Service stations. The latter point-scale observations were interpolated at a 25 km spatial resolution using external drift kriging, wherein the terrain elevation was used as a drift (Samaniego et al., 2013).

2.4 Methods

2.4.1 Selection of Extreme-Meteorological Indices

The ability of the RCMs to reproduce extreme meteorological statistics is evaluated using the indices shown in Table 2.2. These indices are a subset of those used by Sillmann and Roeckner (2008) for the evaluation of Global Climate Models projections, with the exception of $MVal$, which was added to quantify the performance for mean temperature. The precipitation indices presented in Table 2.2 have also been applied for evaluating the ENSEMBLES RCMs for the Iberian peninsula (Herrera et al., 2010; Soares et al., 2012). The main reason for the selection of these extreme indices was their potential impacts (direct or indirect) on the hydrological cycle. For example, the maximum daily maximum temperature (TXx) and the minimum daily minimum temperature (TNn) were selected to display the range of extreme temperature characteristics. Moreover, this temperature range is employed to calculate potential evapotranspiration by the Hargreaves-Samani equation in hydrologic modeling (Samaniego et al., 2010). The number of frozen days (FD) and summer days (SDa) are more representative for processes such as snow pack and actual evapotranspiration, respectively.

ID	Index Name	Index Definition	Unit
Mval	Mean value	Mean value of daily average temperature	°C
TXx	Max Tmax	The maximum daily maximum temperature in each year	°C
TNn	Min Tmin	The minimum daily minimum temperature in each year	°C
FD	Frost days	Mean number of days where minimum temperature is less than 0°C in each year	days
SDa	Summer days	Number of days, where maximum temperature is higher than 25°C in each year	days
RAnn	Annual totals	Annual total precipitation	mm
RX5	Max 5-day precipitation	The maximum of the 5-day precipitation amounts	mm
R95p	Very wet days	The amount of precipitation higher than the 95 percentile of daily precipitation on wet days	mm
CDD	Consecutive dry days	The maximum number of consecutive days with no more than 1mm rainfall	days

Table 2.2: Temperature (upper half) and precipitation (lower half) indices incorporated in this study.

Regarding precipitation, indices such as the five-day maximum precipitation amount ($Rx5$) or the amount of rainfall above the 95th percentile ($R95p$) of daily precipitation were estimated because they are potential triggers of high flows. $Rx5$ is sensitive to the correct representation of the temporal structure of strong precipitation events, while $R95p$ is an indicator for the goodness of fit of the heavy tail of precipitation distribution function. Likewise, the maximum number of consecutive dry days (CDD) is likely to be related to the development of drought events (Samaniego and Bárdossy, 2007). The long-term mean annual total precipitation ($RAnn$), on the contrary, is relevant from the perspective of water resources management.

In general, these indices are calculated on an annual basis for the water year starting on the first of November. This implies a sample size of 39 water years from November 1961 to October 2000. $Rx5$ and $R95p$ are evaluated for the summer and winter water halves of the year separately in order to investigate the characteristics of precipitation extremes during different seasons. These indices are marked by the suffix sum and win in Table 2.3 and Figs. 2.2 and 2.4 for summer and winter, respectively.

2.4.2 The Test Statistic and Hypothesis Testing

The purpose of a test statistic is to condense the entire dataset into a single number for the purpose of performing hypothesis testing. In the present study, the simplest test was to evaluate how the long-term mean (denoted hereafter with the overbar) of the j extreme characteristics simulated by the RCM u at the grid cell i is different from the long-term mean of the corresponding statistic derived from observations. The difference between these two long-term means is called bias. In this study, however, the relative bias B is employed because it normalizes the bias with respect to the observations. This normalization allows to compare model performance among several indices. The relative bias is estimated by

$$B_{iju} = \frac{\overline{\hat{x}_{iju}} - \overline{x_{ij}}}{\overline{x}_{ij}} \,, \tag{2.1}$$

where $\hat{x}$ and x denote the simulated and observed value of a given index, respectively.

The null hypothesis (Ho) in this case is that the simulated value of the index j estimated by the model u at a given location i and the corresponding observation have been drawn from the same distribution. To test this null hypothesis, the two-sided Wilcoxon-Mann-Whitney rank-sum test (Wilks, 2011) is employed because the sampling distribution of B_{iju} is unknown and the underlying variables x are grossly non-Gaussian. This hypothesis test is performed without assuming any asymptotic approximation for the sampling. If this hypothesis is true, then B_{iju} is equal to zero. By resampling the null distribution $\rho = 10,000$ times, the Monte Carlo p-value p_{iju} is estimated. It denotes the probability of obtaining a rank-sum value at least as large or small as the absolute observed value. If p_{iju} is greater than a given significance level α (e.g., 5%), then the null hypothesis would be rejected.

2.4.3 Overall Ranking

The overall score of an RCM is a useful statistic to rank models according to their respective performances. To evaluate the overall score of the RCMs, the absolute values of the relative bias estimated by all models are normalized between zero and one for every index and location. The overall score S for RCM u is then estimated as the arithmetic mean of the normalized biases. Consequently, S_u is calculated as

$$S_u = \frac{1}{NV} \sum_{j=1}^{V} \sum_{i=1}^{N} B^{\eta}_{iju}, \tag{2.2}$$

where V denotes the number of selected indices, and N is the number of cells within the domain. B^{η} is the normalized value of the absolute relative bias obtained with eq. 2.1. A low S_u value indicates a better performance of the RCM u with respect to the other RCMs.

2.4.4 Rejection Rate

In contrast to S, the rejection rate R is an indicator for the statistical robustness of the extreme indices derived from the RCMs. For a given grid cell i and model u, it is estimated by

$$R_{iu} = \frac{1}{V} \sum_{j=1}^{V} 1(p_{iju} < \alpha) \ , \tag{2.3}$$

where $1(p_{iju} < \alpha)$ is equal to one if the p_{iju} is lower than α and otherwise it is zero. R_{iu} thus counts how many indices at location i estimated by the RCM's u have not been drawn from the same distribution as the observations at significance level α. The spatial average $\overline{R}_u$ of the rejection rate can also be used to evaluate and rank the models. The lower the value of $\overline{R}_u$ is, the better the model performance.

2.4.5 Robust Ensemble Selection

It has been frequently argued in the literature that the ensemble average (i.e., mean or median) is a better estimator than a result obtained with a single model, due to the large uncertainty inherent to climate models (Palmer et al., 2005; Tebaldi and Knutti, 2007; Rauscher et al., 2010; Herrera et al., 2010; Sillmann et al., 2013). This study addresses the question of whether a subset of given models can combine the information better than the full ensemble. To address this question, the full ensemble space, considering all possible combinations of models, is evaluated. Five different selection methods are proposed to identify the best-performing combination. Four of them employ an arithmetic mean for the ensemble averaging. The third method is based on the "Reliability Ensemble Averaging" approach (Giorgi and Mearns, 2002) to compare whether a weighted averaging scheme yields a better performance than the other four methods. The mean rejection rate $\overline{R}$ is selected as a performance criterion, which is estimated by the Monte-Carlo p-value for each ensemble average. The estimation of these p-values is, however, computationally costly for the full ensemble space. Therefore, selection methods that evaluate $\overline{R}$ as seldom as possible are required. The proposed methods are the following:

Method 1: Exhaustive ensemble search. This ensemble selection considers all model combinations given U RCMs and selects the combination with the lowest mean rejection rate $\overline{R}$. The total number of combinations is $2^U - 1$. This method is advantageous for a small ensemble because the best combination can be found explicitly. If U is large, perhaps greater than 20, the number of combinations becomes immense and thus too costly in terms of the computational effort required to evaluate $\overline{R}$ for all of the combinations. Therefore, the following four procedures were proposed, which do not require evaluating $\overline{R}$ for all of the combinations. The time complexity of this algorithm is consequently $\mathcal{O}(2^U)$.

Method 2: Score-based ensemble selection. This method ranks and then selects models according to their S values (eq. 2.2) so that, for example, the five top-ranking models constitute the ensemble of size five in this case. The mean rejection rate $\overline{R}$ is then calculated for the average of such subensemble selections. The subensemble exhibiting the lowest $\overline{R}$ among them is then chosen as the most robust combination. This method is the easiest to estimate and therefore it is considered as the benchmark for this analysis. The premise of this selection method is that high performing single models will also constitute a high performing ensemble. Based on this notion, Herrera et al. (2010) chose a subset of five out of nine models that are highly correlated with observations. This method, however, is based on averages that are equally weighted, which motivates the following method.

Method 3: Weighted average ensemble selection. The difference between this method and the other ones is that a weighted ensemble averaging is performed. The weighting scheme is a modification of the "Reliability Ensemble Averaging" (REA) method (Giorgi and Mearns, 2002). A detailed summary of the REA method can be found in Smith et al. (2009). The weights derived by this method are based on "performance" during past periods and "convergence" during future ones. In this study, however, only the performance of models for past periods is taken into account. The weights for the index j and model u are calculated as the average over all locations i by

$$w_{ju} = \frac{1}{N}\sum_{i=1}^{N} \min\left[\frac{\epsilon_{ij}}{B_{iju}}, 1\right], \tag{2.4}$$

where B_{iju} is the relative bias (eq. 2.1) and ϵ_{ij} is the observed variability for the index j at location i. The observed variability ϵ_{ij} is the difference between the minimum and maximum values of a moving average of the observations. Thus, it strongly depends on the size of the moving average window. Choosing a small window (e.g., ten years) leads to high observed variability. This implies that all weights are close to one and the weighted average becomes essentially equally weighted. Giorgi and Mearns (2002) employed a thirty year moving window for a 140-year period. For this study, window sizes between ten and thirty years were tested because only 39 years of data are available. The best results, with respect to the distribution of the weights, were obtained with a 25-year window.

The weighted average for the index j at the ith location and timestep t is calculated as

$$\bar{\hat{x}}_{ij}(t) = \frac{1}{\sum\limits_{u=1}^{U} w_{ju}} \sum_{u=1}^{U} w_{ju}\hat{x}_{iju}(t), \tag{2.5}$$

This ensemble average is then used to estimate the bias (eq. 2.1) and the rejection rate (eq. 2.3). The ensemble average for a subset of models can also be estimated by this equation by setting w_{ju} equal to zero for those models not included in the subset.

The models are then ranked and selected according to their mean weights, i.e., the five models with the highest weights constitute the ensemble of size five. The weights are related to the overall scores S through the relative biases (eqs. 2.4 and 2.2). Thus, the ensembles selected by methods 2 and 3 are the same. The difference stems from the fact that in method 2 all of the weights are equal to $\frac{1}{U}$. The mean rejection rate $\overline{R}$ is then calculated for each weighted ensemble average. The ensemble exhibiting the lowest $\overline{R}$ is then selected as the most robust combination. Methods 2 and 3, however, select the subensembles only based on the performance of single models, which is not the case for the following methods.

Method 4: Forward selection ensemble search. This method resembles the notion of the step-wise regression procedure. As in method 1, the performance of the selected combinations is evaluated using the mean rejection rate $\overline{R}$. The selection procedure is as follows:

1. Evaluate the mean rejection rate $\overline{R}$ over the study domain for all combinations of two models given U ensemble members.
2. Select the combination with the lowest $\overline{R}$ as the “ensemble seed”.
3. Sequentially add the remaining members to the “ensemble seed” and evaluate the corresponding $\overline{R}$.
4. Repeat step 3 for all remaining ensemble members.
5. Replace the old “ensemble seed” with the combination exhibiting the lowest $\overline{R}$ found in steps 3 and 4.
6. Repeat steps 3 to 5 until all ensemble members are included.
7. The combination with the lowest $\overline{R}$ is the most robust ensemble combination.

This method is much more efficient than method 1 because a maximum of $(U - 1)^2$ model combinations have to be evaluated. Nevertheless, half of the computational cost is required for finding the first “ensemble seed”, which is the best combination of two models (step 1). Additionally, the performance of this procedure does depend on the number of models considered for the first “ensemble seed”. This potential drawback is addressed in method 5.

Method 5: Backward elimination ensemble search. As in method 4, this method tries to find the best-performing combinations of models without evaluating the full ensemble space. The difference from method 4 is that it does not start with a small ensemble, but with the largest possible. The exact procedure is as follows:

1. Select the full ensemble of all models as the "ensemble seed".
2. Sequentially remove a remaining member from the "ensemble seed" and evaluate the corresponding $\overline{R}$.
3. Repeat step 2 for all remaining ensemble members.
4. Replace the old "ensemble seed" with the combination exhibiting the lowest $\overline{R}$ found in steps 2 and 3.
5. Repeat steps 2 to 4 until the "ensemble seed" contains only a single model.
6. The combination with the lowest $\overline{R}$ is the most robust ensemble combination.

Compared to method 4, the computational cost is further reduced to $U(U+1)/2-1$. The time complexity of methods 4 and 5 scales with $\mathcal{O}(U^2)$.

Ensemble groups selected by method 2 are denoted by S, e.g., $S5$ denotes the five member ensemble group selected by score-based selection (method 1). Ensemble groups selected by methods 1, 3, 4, and 5 are denoted by O (optimal), W (weighted average), F (forward selection), and B (backward elimination), respectively.

2.5 Results and Discussion

2.5.1 Dimensionality of the Input Data

Based on the PCA analysis (Wilks, 2011) of the precipitation data sets (i.e., EOBS and REGNIE), EOBS exhibits less spatial variability than REGNIE, irrespective of the season. This is because the number of principal components required to explain a given percentage of variance (i.e., dimensionality) in EOBS is always less than that required for REGNIE (Fig. 2.1). In other words, the dimensionality of REGNIE is higher than that of EOBS, and thus it can be regarded as the best reference data set available for the evaluation of RCM precipitation and derived statistics over Germany.

Although the products were compared at the same spatial resolution (25 km), the difference in dimensionality stems from the fact that the REGNIE data set incorporates information from a relatively denser meteorologic station network, which is at least twice as large as that of EOBS. In addition, the regionalization technique employed for the spatial interpolation considered the stochastic dependency of precipitation on landscape features such as terrain elevation and aspect.

The difference in dimensionality of the air temperature fields obtained with interpolated data and EOBS was not significant ($< 1\%$ on average). Nevertheless, to be consistent with the previous selection, the interpolated data set, which takes into account the stochastic dependency between temperature and elevation, was used as a reference for further analysis.

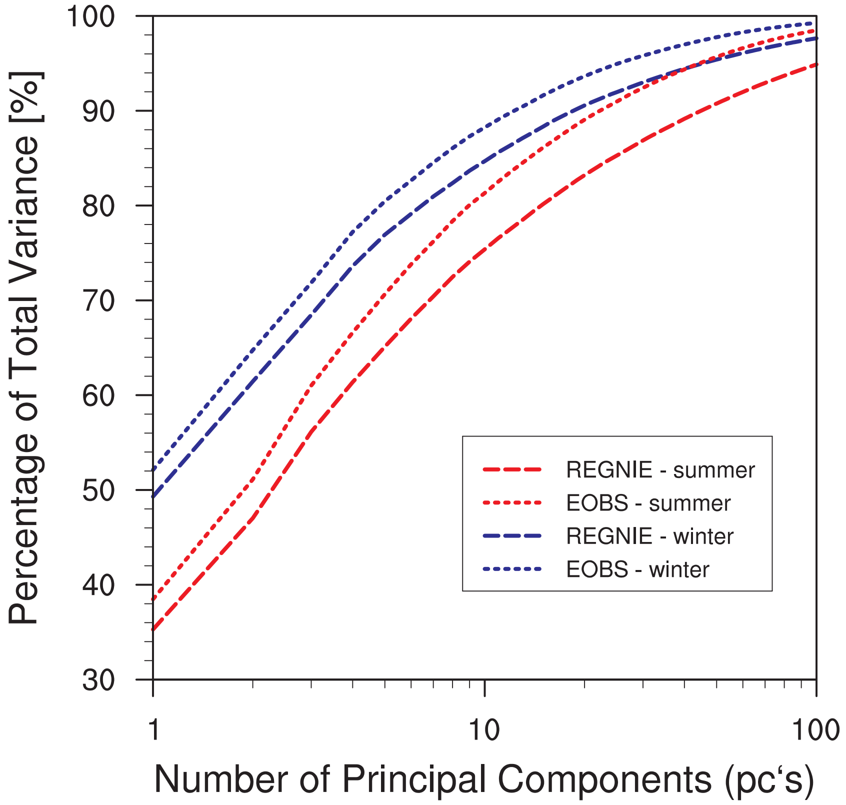


Figure 2.1: Percentage of total variance explained by a given number of principal components for the aggregated REGNIE product (red lines) and the EOBS data set (blue lines) for two periods: summer (red lines) and winter (blue lines).

2.5.2 Statistics and Tests

The range of the relative biases of the single RCMs are shown in Fig. 2.2a for each temperature index (Table 2.2). Indices TXx and $Mval$ exhibit the smallest spread, ranging from -20% to +10%. The spreads of the indices TNn, FD, and SDa increase gradually up to a range of -100% to +40%.

The relative biases obtained for $Mval$ are evenly distributed within ±10%, with the exception of model K. This is the narrowest spread for all indices. This could have been expected because climate models are mainly calibrated against mean annual temperature. The ENSEMBLES RCMs are equally good at representing TXx. Half of the models also exhibit a positive or negative bias smaller than 10%, with the exception of model M.

Models M and K are the only models that drastically underestimate TNn, up to a value of -40%. The majority of the models overestimate TNn but not exceeding a relative bias of 20%. As a consequence, an ensemble average of all models will also overestimate this index. This, however, might not be the case for a chosen subset of models. Although TNn is overestimated by most models, there is no tendency to underestimate FD. For example, model C overestimates TNn by +20%, but exhibits almost no bias for FD. Therefore, freezing periods are generally well captured in extent by this

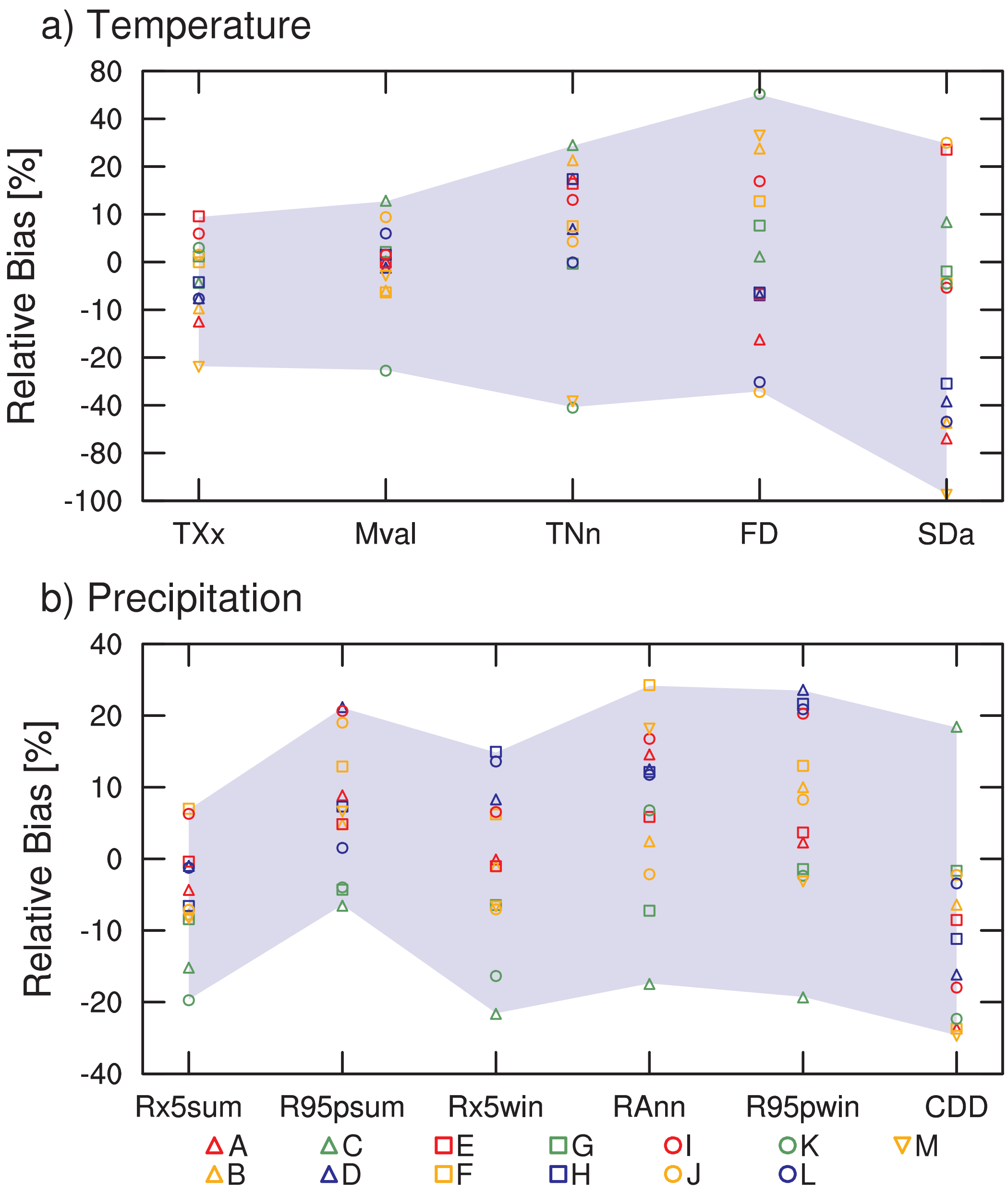


Figure 2.2: Spatial average of relative bias B for every index and RCM. Panels a) and b) depict the performance of precipitation and temperature indices, respectively. The statistical summary of theses indices is shown in Table 2.3.

model, but they are too warm. The biases in FD and SDa are more strongly pronounced than in the other temperature indices. This stems from the fact that both indices are related to the explicit thresholds of $0°$C and $25°$C, respectively. If an RCM is not able to exceed or undercut this threshold, then the model automatically exhibits a stronger bias. Model M, for example, has a long-term average for TXx of approximately $24°$C (Table 2.3), which does not exceed $25°$C, and therefore it yields a

Temperature

Abbreviation	Mval	TXx	TNn	FD	SDa	Score_Tem	Rank
Observed	8.49	31.79	−14.59	88.46	28.04		
A	8.50	27.86***	−12.09	74.36	8.91 ***	0.377	10
B	7.98	28.70***	−11.35**	112.20 ***	12.76 ***	0.404	11
C	9.57***	30.40*	−10.47***	89.47	30.39	0.319	8
D	8.39	29.37***	−13.58	82.30	17.43 **	0.193	3
E	8.48	34.84***	−12.24*	82.49	35.44 *	0.235	6
F	7.95	31.78	−13.49	99.65	26.80	0.141	2
G	8.67	32.17	−14.64	95.23	27.49	0.089	1
H	8.62	30.45	−12.10*	82.80	19.61 *	0.198	5
I	8.62	33.69	−12.71	103.17 *	26.53	0.196	4
J	9.29**	32.28	−13.96	58.80 ***	36.23 *	0.293	7
K	6.38***	32.74	−20.68***	141.84 ***	26.77	0.636	12
L	9.00*	29.35**	−14.60	62.40 ***	13.06 ***	0.328	9
M	8.25	24.37***	−20.10**	116.93 ***	0.69 ***	0.699	13
13	8.44	30.62	−14.00	92.43	21.70	0.103	
O9	8.56	31.70	−13.89	88.72	25.88	0.053	
S9	8.56	31.40	−13.18	86.56	23.93	0.087	
W9	8.54	31.61	−13.34	88.45	25.83	0.068	
F9	8.51	31.97	−13.88	90.65	26.92	0.054	
B9	8.56	31.70	−13.89	88.72	25.88	0.053	

Precipitation

Abbreviation	RAnn	Rx5sum	Rx5win	R95psum	R95pwin	CDD	Score_Pr	Rank
Observed	775.73	56.32	48.84	127.22	106.22	21.25		
A	883.70**	53.78	48.79	138.63	108.76	15.60**	0.443	6
B	795.60	51.85	48.32	133.59	116.85	19.85	0.295	3
C	645.76**	48.17	37.75*	118.66	85.97	25.05	0.467	8
D	869.936∗	55.74	52.96	155.09	134.27	17.98	0.505	11
E	822.71*	56.10	48.32	133.63	110.30	19.41	0.289	2
F	991.09***	60.37	51.97	143.08	119.55	15.59**	0.564	13
G	718.19	51.50	45.60	121.53	104.65	20.88	0.251	1
H	867.45**	52.51	55.80	136.73	130.35	18.91	0.494	10
I	899.85**	59.97	52.13	153.96	127.91	17.57*	0.508	12
J	758.30	52.21	45.31	150.93	115.16	20.75	0.390	5
K	829.80**	45.27*	41.23	121.95	103.60	16.16**	0.489	9
L	864.84*	55.59	55.22	129.26	128.96	20.49	0.343	4
M	912.15***	51.57	45.51	135.87	102.73	15.16**	0.456	7
13	835.34*	53.43	48.38	136.38	114.54	18.72	0.272	
O9	808.54	53.01	48.26	135.75	115.69	19.69	0.241	
S9	843.11	55.09	50.62	139.76	120.89	19.05	0.301	
W9	823.64*	55.07	50.40	137.98	119.95	19.47	0.271	
F9	822.00	53.52	48.15	134.41	114.05	19.42	0.245	
B9	808.54	53.01	48.26	135.75	115.69	19.69	0.241	

Table 2.3: Spatial average of long term values for each Temperature and Precipitation index for the observations, all single models, and six ensembles. These are the full ensemble (13), the optimal ensemble of size nine (O9), the score based ensemble of size nine (S9), the weighted ensemble average of size nine (W9), the forward selected ensemble of size nine (F9), and the backward selected ensemble of size nine (B9). * indicates that the null hypothesis is rejected in 50% of the cell with the 5% significance level, where as ** and *** indicate that 50% of the cells are significant at the 1% and 0.1% significance level, respectively. Additionally the aggregated scores for Precipitation and Temperature as well as the ranking of these are shown.

large bias in SDa. SDa is underestimated by most models, but it is especially strong by those that also underestimate TXx, such as model A.

The spatial pattern of the observed long-term mean for each temperature index is shown in the first column in Fig. 2.3. The second, third, and fourth columns depict the spatial patterns shown by the worst and the best models and the full ensemble average for each index. Distinctive patterns can be seen for the observed indices. For example, the pattern of $Mval$ shows that the warmest regions in Germany are located in west and south-west Germany. Lower $Mval$ values can be observed in central Germany and the prealpine regions that are characterized by mountainous terrain. The patterns of TNn and $Mval$ are directly related, whereas those of FD and $Mval$ are inversely related. The patterns of TXx and SDa are comparable. Both exhibit high values in south-west and central-east Germany.

Models A and F are the best models for $Mval$ and TXx, respectively. They exhibit a bias that is in general less than 5% in magnitude over the entire domain (Fig. 2.3). Therefore, their values realistically resemble the observed patterns. The ensemble average of all models also shares this characteristic, which could have been expected because the biases are evenly distributed around zero (Fig. 2.2a).

Models L, D, and F are the best-performing models for TNn, FD, and SDa, respectively. All models show a distinctive pattern exhibiting pronounced over- and underestimation up to values of $\pm30\%$. The models tend to overestimate in regions exhibiting lower observed values and underestimate in regions exhibiting higher observed values (Fig. 2.3). For example, model L underestimates TNn in central-west Germany, whereas model D overestimates FD in this region. Model F underestimates SDa in central-south Germany but overestimates it in coastal areas in the north. This behavior implies that the RCMs are not able to correctly represent the observed spatial variability of these indices. In other words, the simulated fields have too little spatial variability. Although the biases are less pronounced for the full ensemble average, this behavior is still present for the indices TNn and FD. In contrast, the full ensemble average exhibits a strong underestimation of SDa. This is the only index for which the bias of the full ensemble average is greater than that obtained with the single best model (model F for SDa).

The worst models show either a strong over- or underestimation of the observed values for all indices (second column in Fig. 2.3). The biases do not exhibit any distinctive spatial pattern. Only models K and M are selected as the worst models for all indices. Therefore, a subensemble that does not consider these two models would exhibit less bias than the full ensemble.

The range of the relative biases of the single RCMs are shown in Fig. 2.2b for each precipitation index listed in Table 2.2. As shown in this figure, the relative biases for $RAnn$ vary from -20% to +20%. Therefore, the range of modeled precipitation biases is twice as high as that of modeled temperature biases, considering $Mval$ as the reference temperature index (Fig. 2.2a). Ten of the 13 RCMs overestimate $RAnn$, which implies that there is too much precipitating water over the land surface. Most RCMs overestimate $R95psum$ and $R95pwin$, which contribute to the overestimation of $RAnn$. This implies that a subensemble average could provide a less biased estimate than the full ensemble average.

The relative biases for $Rx5win$ exhibit the most evenly distributed values around zero. $Rx5sum$, on the contrary, is underestimated by most models. This is an indication that the mechanisms for

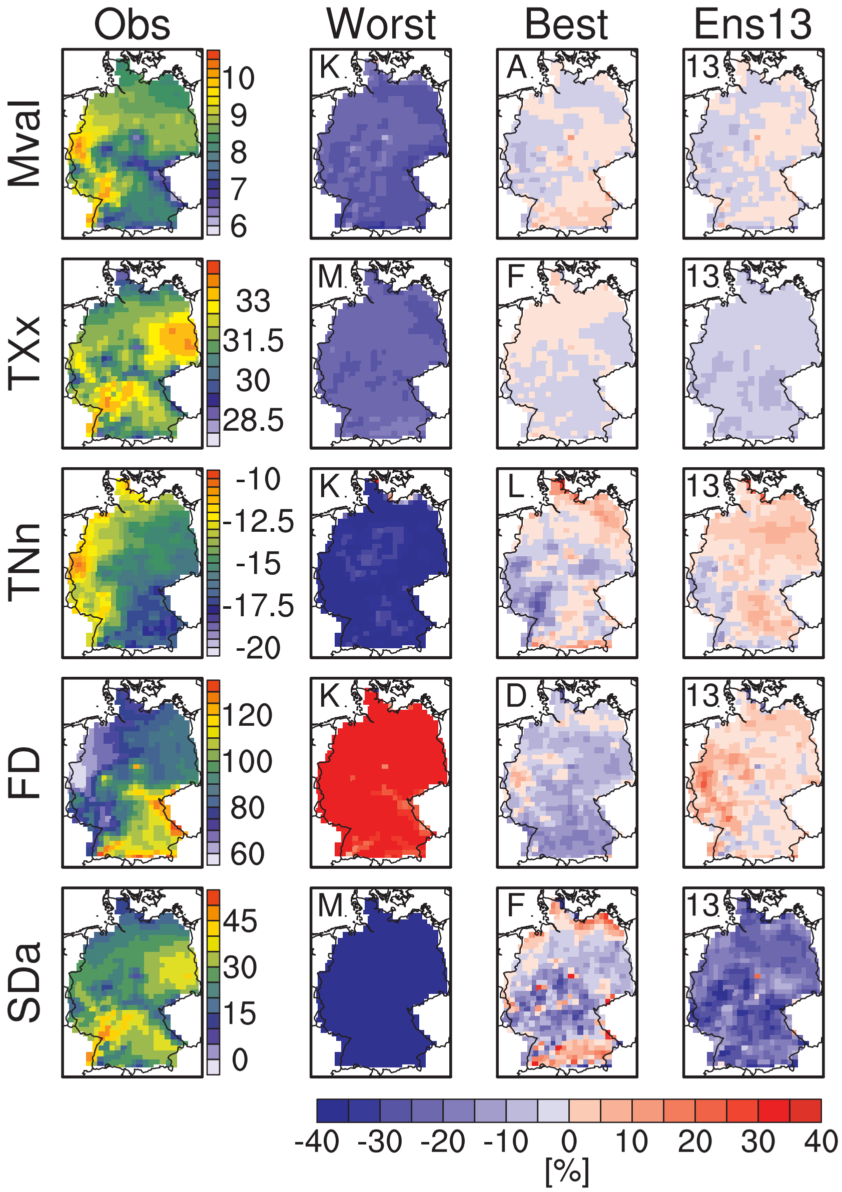


Figure 2.3: Spatial Pattern for each temperature index (rows). The spatial pattern for the observations is shown in the first column. The units for each index are encompassed in Table 2.2. The second, third, and fourth column show the relative bias in percent for the worst performing model for each index, the best performing model for each index, and the ensemble average of all models. The worst- and best-performing RCMs are assessed with respect to the interpolated observational data set.

generating precipitation in the RCMs allocate too little precipitation to consecutive extreme wet days during summer, while annual precipitation is generally overestimated.

CDD is the only index measuring dry events. This index is underestimated by all models except model C (Fig. 2.2b). A similar behavior has been reported for Spain (Herrera et al., 2010). This implies that, regardless of the climatic zone, the precipitation generation mechanisms are consistently misrepresented.

The spatial pattern of the observed long-term mean for each precipitation index is shown in the first column of Fig. 2.4. The second, third, and fourth columns depict the spatial pattern shown by the worst and best model for each index and the ensemble average of all models. The patterns for the observed precipitation indices are strongly influenced by the German orography, with lower values in north-east Germany and higher ones over mountainous regions in central and southern Germany. The pattern for CDD is inversely related to the pattern for the other indices.

Models G, E, M, L, and M are the best models for the indices $RAnn$, $Rx5sum$, $Rx5win$, $R95psum$, and $R95pwin$, respectively (third column). All of these models overestimate precipitation in north-east Germany, where the observed values are generally low. Underestimation occurs mostly in regions of higher observed values. Consequently, RCMs lack the capability to represent the observed spatial variability of precipitation over Germany, which was also observed for temperature indices. The biases for the full ensemble average (fourth column) are more pronounced compared to the bias of the best-performing models (third column) for the indices $RAnn$, $Rx5win$, $R95psum$, and $R95pwin$. This is a clear indication that there are models that detract heavily from the ensemble performance.

The worst-performing models (second column) tend to show either more overestimation or more underestimation over the whole domain for all indices. Only model D, which is the worst-performing model for $R95psum$ exhibits both, over- and underestimation. This model overestimates $R95psum$ over most regions in Germany, with the exception of the prealpine. The best-performing model L and the full ensemble average also underestimate in this region. Therefore, it can be concluded that most models are underestimating extreme rainfall events in this mountainous region during summer.

CDD is underestimated by most models (Fig. 2.2b). Therefore, the full ensemble average also underestimated this index over the whole domain (Fig. 2.4). This again indicates that a subensemble average would perform better than the full ensemble average. However, it can be seen that there are six different worst-performing models, depending on the chosen precipitation index (second column in Fig. 2.4). As a result, a high performing subensemble might not be found.

In general, RCMs exhibit substantial biases for both temperature and precipitation indices (Fig. 2.2), either consistently over- or underestimating specific indices. For example, CDD is underestimated by most models (Fig. 2.2b), which has also been reported for Spain (Herrera et al., 2010). Furthermore, over- and underestimation are strongly related to regions with lower and higher observed values, respectively. Therefore, the analyzed RCMs lack the ability to represent the observed spatial variability of these indices. The full ensemble is also biased for indices that are consistently over- or underestimated by all models (e.g., TNn or CDD). Investigating weighted ensemble averaging and subensemble selection methods could provide a worthwhile alternative.

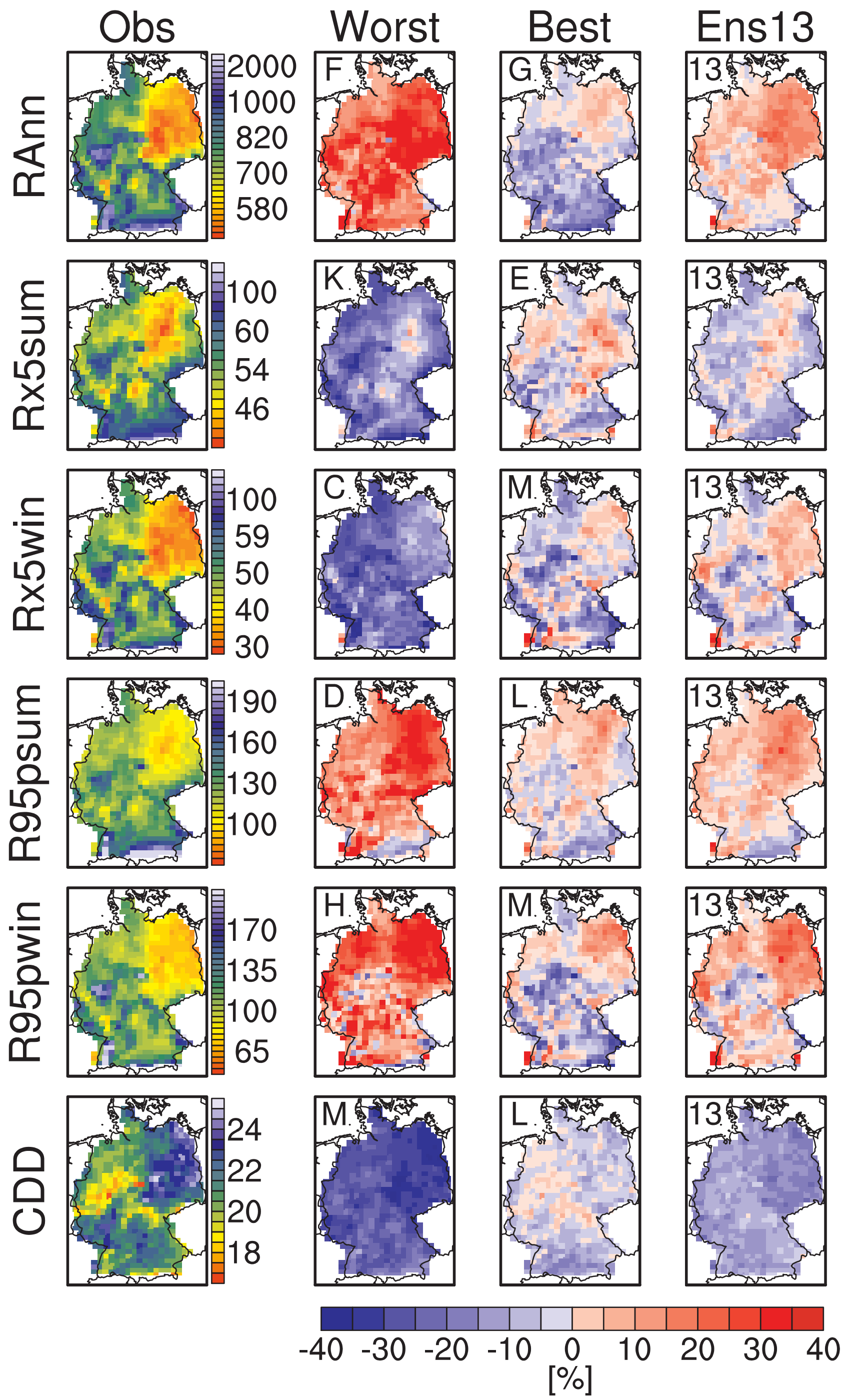


Figure 2.4: Analogous to Fig. 2.3, only for precipitation indices. The worst- and best-performing RCMs are assessed with respect to the REGNIE observational data set.

2.5.3 Reliability Ensemble Averaging

Method 3 as described in Section 2.4.5 was applied to investigate the performance of the weighted ensemble averaging. The derived weights for each index and RCM are depicted in Fig. 2.5.

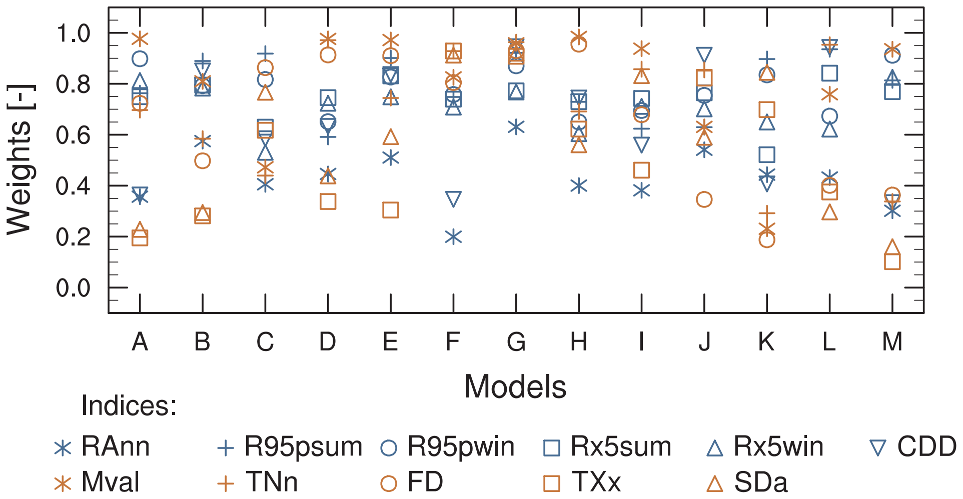


Figure 2.5: Spatially averaged weights for each index derived by the "Reliability Ensemble Averaging" (REA/method 5) for each model. Red and blue markers indicate temperature and precipitation indices, respectively.

Model G was most heavily weighted among all of the models, which means that it exhibits the least bias and is outperforming the other models. This result supports the findings of Soares et al. (2012), who also reported that model G was the best-performing model for Portugal, especially for reproducing precipitation extremes. This implies that model G is able to reproduce extreme precipitation over different climatic zones. The weights of model G and F with respect to temperature indices were slightly higher than those obtained for the precipitation indices, which could have been expected because temperature variability is less complex than precipitation. However, most of the other models do not exhibit a clear tendency in this respect. This can be partially explained by the fact that the observed variability (ϵ in eq.2.4) for the temperature indices is less than that obtained for the precipitation indices. Therefore, for a given relative bias, the estimated weights for temperature indices tend to be lower than those obtained for precipitation indices.

Because model biases and weights are related by eq. 2.4, models exhibiting the highest bias for a given index also exhibit the lowest weight. For example, models K and M exhibit the lowest weights for temperature indices, which correspond to the greatest biases depicted in Fig. 2.3. In contrast, models L and G, which have the smallest bias for $R95psum$ and for $RAnn$, respectively (Fig. 2.4), exhibit the highest weights for those indices.

It can be expected that the weighted ensemble averaging of all models would produce a less biased estimate than the equal-weighted ensemble averaging because higher biased models have less impact on the average. However, it is yet to be investigated if this advantage is still beneficial for subensembles.

2.5.4 Model Ranking and Ensemble Performance

The final score S and the mean rejection rate $\overline{R}$ for each single model and six ensemble combinations are depicted in Figs. 2.6a and 2.6b. Most models have an S value between 0.32 and 0.42, with the exceptions of E, G, K, and M (Fig. 2.6a), which can be considered as outliers. Models E and G are high performing models exhibiting S values of 0.18 and 0.26, respectively, whereas models K and M are poor performing models exhibiting S values of 0.55 and 0.57, respectively. The final scores for the considered ensembles cluster between 0.15 and 0.20. Only the combinations found by methods 1, 4, and 5 (O9, F9, and B9, respectively) are able to outperform the single best-performing model G.

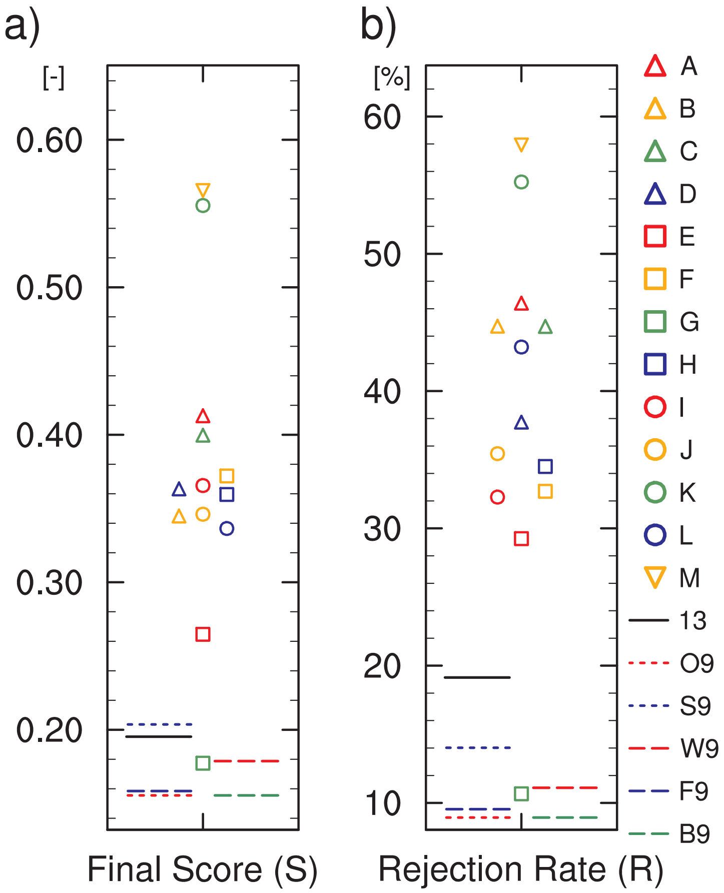


Figure 2.6: Final score S (panel a) and rejection rate $\overline{R}$ (panel b) for single models and selected subensembles identified by different methods. Different search methods are indicated by: O - method 1, S - method 2, W - method 3, F - method 4, and B - method 5. For clarity, the x-axis position of the markers are varied.

The mean rejection rate $\overline{R}$ varies from 29% to 47%, with the exception of models G, K, and M (Fig. 2.6b). Only 11% of the indices are rejected for model G, which highlights the superior performance of this model. Models K and M are again the worst-performing models, exhibiting rejection rates of 55% and 58%, respectively. Although the ordering of the worst two and best two models is the same for both metrics, this is not usually the case. For example, model L is the third best-performing model with regard to S (Fig. 2.6a), but it is also the sixth worst-performing model with regard to $\overline{R}$ (Fig. 2.6b). Consequently, it is not necessarily true that models exhibiting a relatively low bias would also exhibit a low rejection rate. This result shows the added value of evaluating both the magnitude and statistical significance of the bias.

The only ensembles that are able to outperform the single best-performing model G are the subensembles O9, F9, and B9. The difference in the rejection rate between them and the full ensemble denoted as “13” in Fig. 2.6b is almost 10%, which confirms the hypothesis that some subensembles can reproduce the observations more faithfully than the full ensemble.

The full spread of the mean rejection rate $\overline{R}$ for all possible combinations of models is shown in Fig. 2.7a. Overall, $\overline{R}$ varies from 10% to 70%. The range is largest for combinations with fewer than four members and narrows with increasing ensemble size. This should be expected because the impact of a single model on the ensemble average decreases as the amount of information increases. It is interesting to note that the median $\overline{R}$ decreases significantly up to ensembles of size seven and then remains almost constant for larger ensembles. This indicates that for the given 13 RCMs, it is in general advantageous to consider the ensemble averages instead of single models.

Additionally, Fig. 2.7a depicts the performance of the ensemble selection methods presented in Section 2.4.5. The combinations found by method 1 are the best possible ones for a given ensemble size. These optimal combinations are indicated by the corresponding lower bar in Fig. 2.7a. They exhibit a constant $\overline{R}$ of approximately 10% for ensemble sizes smaller than nine. The $\overline{R}$ value then gradually increases for larger ensemble sizes up to 20%, which corresponds to the value obtained for the full ensemble average. In this study, 8,191 evaluations of $\overline{R}$ were carried out by this method.

The combinations found by method 2 (black dots in Fig. 2.7a) exhibit higher $\overline{R}$ values than those obtained by method 1 with the exception of ensemble sizes 1, 12, and 13. Model G would be selected by method 1 as the best subensemble because it has the lowest $\overline{R}$ among all combinations considered by this method. This method required only 13 evaluations of $\overline{R}$.

The combinations selected by method 3 are the same as those of method 2, although weighted ensemble averaging is employed. The $\overline{R}$ values for these combinations vary from 10% to 15% and exhibit no tendency with respect to ensemble size. This is because highly biased models are weighted the least. Although the mean rejection rate for these combinations is generally low, they are not as good as those obtained by method 1 for ensembles of six to nine models. Therefore, the weighted ensemble averaging method does not fully utilize the information contained in the RCMs. This method also required only 13 evaluations of $\overline{R}$.

The combinations found by method 4 (yellow dots in Fig. 2.7a) exhibit $\overline{R}$ values that are closer to the minimum than those obtained by method 2, but they are not optimal. This is because the ensemble average for small ensemble sizes strongly depends on selected single models. The optimal combinations for ensemble sizes of two, three, and four contain different models, with only a single model in common. The forward selection search method (method 4) sequentially adds models to a combination that has already been found. Therefore, it is not able to correctly identify the best combinations for these ensemble sizes. This method required 144 evaluations of $\overline{R}$.

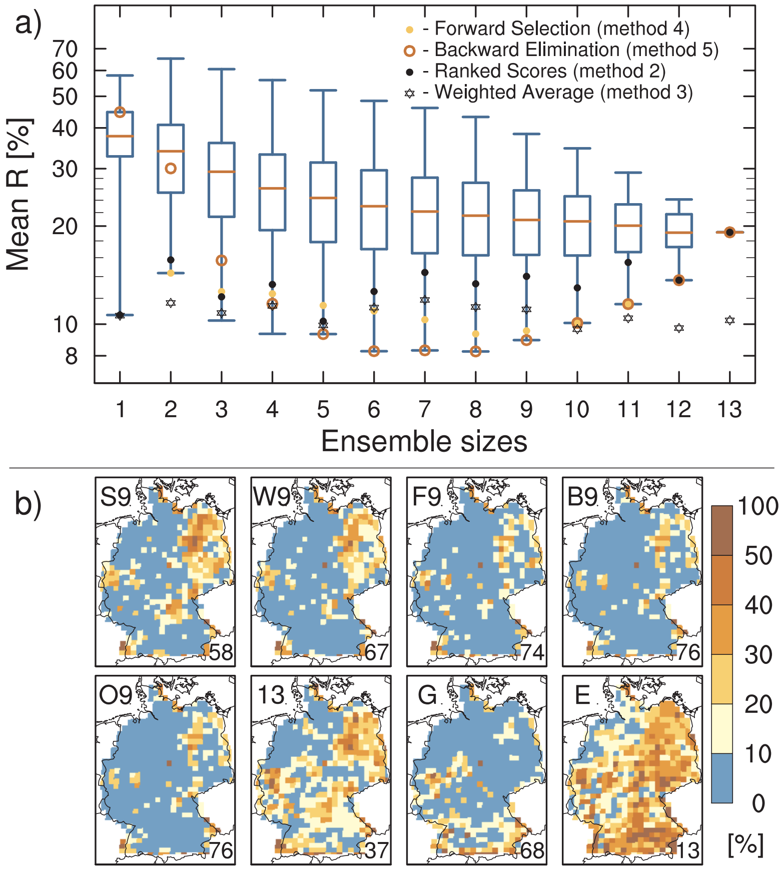


Figure 2.7: Top panel a): Variability of the rejection rate $\overline{R}$ for all ensembles of different sizes is shown with box-plots. The box-plots range from minimum to maximum $\overline{R}$ value with 50% of the $\overline{R}$ values between the 25th and 75th percentile in the box. The median is marked as red bar. Furthermore, the performance of the selection methods is shown. Yellow dots represent method 4, red circles method 5, black dots method 2, and black stars method 3. The $\overline{R}$ of the subensemble selected by method 1 is marked by the lower bar. For clarity, a logarithmic y-axis is used. Bottom panel b): The spatial performance for ensembles of nine models selected according to method 2, 3, 4, and 5 (i.e. S9, W9, F9, and B9, respectively) is shown in the first row. The spatial performance for the ensemble of nine models selected by method 1 (O9), the full ensemble (13), and two single models (the best - G and the second best - E) is shown in the second row. The rejection rate was estimated at the 5% significance level.

The combinations found by method 5 (red circles in Fig. 2.7a), on the other hand, exhibit a close match with those combinations found by method 1 for ensemble sizes larger than four. For ensemble sizes smaller than five, the combinations obtained by this method are not comparable to the best combinations obtained by method 1 because they have only one model in common. Therefore, because this method sequentially removes models from a combination that has already been found, it is not able to correctly identify subensembles with small sample sizes. This method, however, is able to identify the best combinations for ensemble sizes larger than eight. Moreover, it is quite advantageous compared with method 1 because it only required 90 evaluations of $\overline{R}$ in this example, which is only 1% of the evaluations required by method 1.

The spatial distribution of the rejection rate R for combinations of nine models identified by each selection method are shown in Fig. 2.7b. The spatial distribution for the full ensemble average and the two best-performing single models G and E is also depicted. The ensembles of nine models were selected because the $\overline{R}$ values gradually increase for the larger ensembles selected by methods 1, 2, 4, and 5 (Fig. 2.7a). The number of cells exhibiting an R less than 10% (blue cells) shown in Fig. 2.7b is higher for all combinations identified by any selection method compared to the full ensemble. The subensembles S9, W9, F9, B9, and O9 shown in Fig. 2.7b have lower rejection rates than the full ensemble (13) over entire regions, for example, in west, central-south, and east Germany. Additionally, within the combinations found by the single selection methods, the subensembles selected by methods 1, 4, and 5 have lower rejection rates than those found by methods 2 and 3, in particular in east and west Germany.

The single best-performing model G exhibits a relatively low rejection rate, as seen by the high percentage of blue cells (68%). The second best-performing model E only exhibits 13% blue cells. This means that this model fails to reproduce observed indices over 87% of the territory.

Although these are promising results, the analysis was conducted in a situation where one RCM (model G) was outperforming every other model, as shown in Figs. 2.6a and 2.6b. In other situations, this might not be the case. To investigate the effect of model G on the performance of subensembles, the subensemble selection was repeated excluding model G.

The full spread of the mean rejection rate $\overline{R}$ for all possible combinations of models without considering model G is shown in Fig. 2.8. It is analogous to Fig. 2.7 where model G is included. The major difference between the upper panels of these figures is that the minimum $\overline{R}$ value increases to 30% for small sample sizes (less than four) if model G is excluded. The strongest impact occurs for the subensembles identified by method 2 because model G was always selected by this method. The impact on the other selection methods was negligible. The minimum rejection rates are obtained by methods 1 and 5 for ensembles of size five to eight. Both methods select the same subensembles for ensembles larger than five. The spatial distributions of the subensembles (S8, W8, F8, B8, and O8) are comparable to the subensembles depicted in Fig. 2.7b, even though the model G was not included. Therefore, if model G had not been part of the ENSEMBLES project, it would have been highly beneficial to use subensembles instead of single models or the full ensemble.

Overall, subensemble averages are able to reproduce observed extremes more robustly than single RCMs and the full ensemble average. However, only the best subensembles are able to outperform the best single model G. The backward elimination method is capable of identifying these subensembles and only demands 1% of the evaluations of $\overline{R}$ required by method 1.

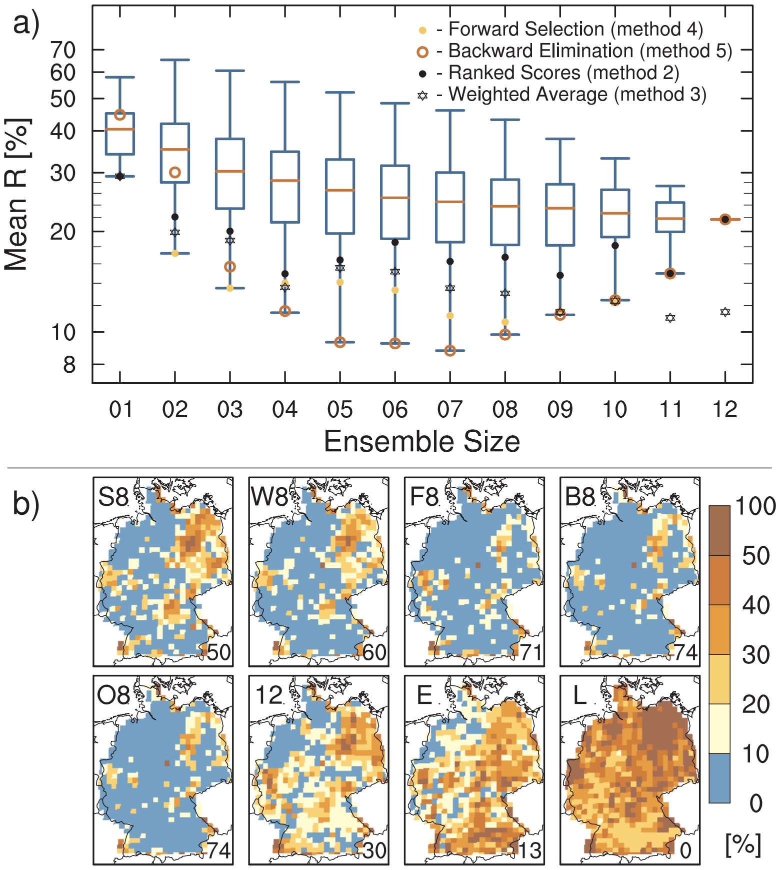


Figure 2.8: Analogous to Fig. 2.7 without considering model G. Bottom panel b): The spatial performance for ensembles of eight models selected according to method 2, 3, 4, and 5 (i.e. S8, W8, F8, and B8, respectively) is shown in the first row. The spatial performance for the ensemble of eight models selected by method 1 (O8), the full ensemble (12), and two single models (the second best - E and the third best - L) is shown in the second row.

2.5.5 Implications for Policy Recommendations and Impact Assessments

The foregoing results could have broad implications for policy recommendations because the ensemble average can be used directly. The minimum $\overline{R}$ values obtained for the subensembles selected by methods 1 and 5 are almost constant for the subensembles with sizes ranging from six to eight (Fig. 2.7a). The spatial pattern for the subensembles of size six (not shown in this figure) is quite similar (rank correlation equal 0.95) to those obtained for subensembles O9 and B9 in Fig. 2.7b. Consequently, six out of 13 RCMs are effectively able to reproduce the best possible ensemble average. Moreover, taking more models into account does not increase the ensemble performance and, therefore, those models may be discarded for studies focusing on policy recommendations.

For hydrologic impact assessments, single model runs are required rather than ensemble averages. It can be assumed that the bias in the meteorological forcing will propagate nonlinearly through the hydrologic impact model. Therefore, the subensemble exhibiting the best ensemble average with regard to meteorological indices might not be necessarily the subensemble that generates the least biased hydrologic realizations. Nevertheless, it can be assumed that if an RCM exhibits a greater bias with regards to a meteorological index that is relevant for hydrology, then the hydrologic variable of interest would also be highly biased. For example, if an RCM overestimates $RAnn$ (e.g., model M by 15%), then variables such as streamflow would also be biased.

Therefore, subensembles should ideally fulfill two conditions if they are to be used for impact assessments and policy recommendations that are based on those assessments. The first condition is that the ensemble average exhibit a mean rejection rate $\overline{R}$ close to the minimum $\overline{R}$. The second condition is that the selected models exhibit a low bias. The subensembles selected by method 1 fulfill the first condition but not necessarily the second. The subensembles selected by method 2 fulfill the second condition but not necessarily the first. Therefore, if methods 1 and 2 select the same subensemble, then both conditions are fulfilled and this subensemble can be used for impact assessments. Because this subensemble is composed of single RCMs that exhibit the closest match to observations and their ensemble average is the best possible one, including additional models in this combination would increase the bias of the ensemble average.

These two conditions, however, are so stringent that they are not fulfilled by any subensemble in the present study. A relaxed version of these conditions may still be useful for impact assessments. For example, method 1 identifies subensembles that fulfill the first condition (i.e., ensembles of size six, seven, and nine. The minimum $\overline{R}$ is obtained for ensemble of size eight). From these subensembles, the one that best fulfills the second condition can be identified (subensemble of size nine). This subensemble can be used for impact assessments for the following reasons.

The minimum rejection rate $\overline{R}$ of subensembles gradually increases from ensembles of nine models by taking more models into account (Fig. 2.7a). The combinations found by method 3 do not share this behavior. However, this method is not useful for impact assessments because the derived weights cannot be directly transferred to the impact model due to its intrinsic nonlinearities. An increase of the rejection rate for large subensembles can only occur through the subsequent addition of strongly biased models. Here, this deterioration from the absolute minimum occurs irrespective of whether model G is considered (Figs. 2.7a and 2.8a). Therefore, the second condition is fulfilled to some extent by the best subensemble of size nine. The deterioration of $\overline{R}$ is caused by including either models A, B, F, or M.

Consequently, RCMs A, B, and M can be discarded for impact assessments because they belong to the worst-performing models in terms of the mean rejection rate $\overline{R}$ (Fig. 2.6b). Taking them into account would only lead to a more biased ensemble average and very likely to more biased hydrologic realizations. Model K also exhibits a high $\overline{R}$ value. However, if this model had been discarded, then the overall ensemble average would have been more biased. In other words, model K is retained in the selected subensemble to balance out the bias of the other selected models. Although model F is a relatively well-performing model for the metrics S and $\overline{R}$ (Fig. 2.6), if it would be included in the subensemble O9, then the rejection rate would increase. Therefore, it should not be discarded but rather retained in the evaluation so that impact assessments can be performed with and without this model.

In general, the different subensembles extract useful information from the RCMs for either policy recommendations or impact assessments because the two applications require different conditions. A subensemble that contains six out of 13 RCMs can be used for policy recommendations. Furthermore, three models have been identified that could be discarded for hydrological impact assessments.

2.6 Summary and Conclusions

In this study, the performances of 13 RCMs from the ENSEMBLES project were analyzed to investigate whether these models are able to reproduce the extreme indices of precipitation and temperature observed in Germany between 1961 and 2000. This study also aimed to find out whether subensembles could provide a better representation of the observed variability for these indices than a single RCM or the full ensemble. The selected indices are highly relevant for hydrologic modeling, water resources impact assessment, and providing policy recommendations. The magnitude and rejection rate of the model bias was evaluated in this study.

ENSEMBLES RCMs excel at reproducing long-term mean annual temperature $Mval$ with biases between $\pm10\%$. The range for precipitation indices is generally $\pm20\%$. The models tend to consistently over- or underestimate observations for specific indices. For example, long-term annual precipitation ($RAnn$) tends to be overestimated, whereas the maximum number of consecutive dry days (CDD) is consistently underestimated. These two results have also been reported by Rauscher et al. (2010) and Herrera et al. (2010), respectively. The models also underestimate the observed spatial variability of extreme index values. The ranking of RCMs based on the rejection rate is different from that obtained from the magnitude of the overall bias. This highlights the added value of performing a test of statistical significance. Model G (KNMI-RACMO2) was able to outperform all of the other models in terms of both bias and rejection rate. This model was also able to excel in Portugal (Soares et al., 2012).

This study has shown that subensemble averages are able to outperform both single models and the full ensemble average. With regard to policy recommendations, six out of 13 RCMs can best estimate the values generated from the observations. For impact assessments, three out of the 13 RCMs drastically impair the ensemble performance and should be discarded. The combinations identified for policy recommendations and impact assessments reduce the rejection rate over several regions in Germany, such as central-south, west, and east Germany.

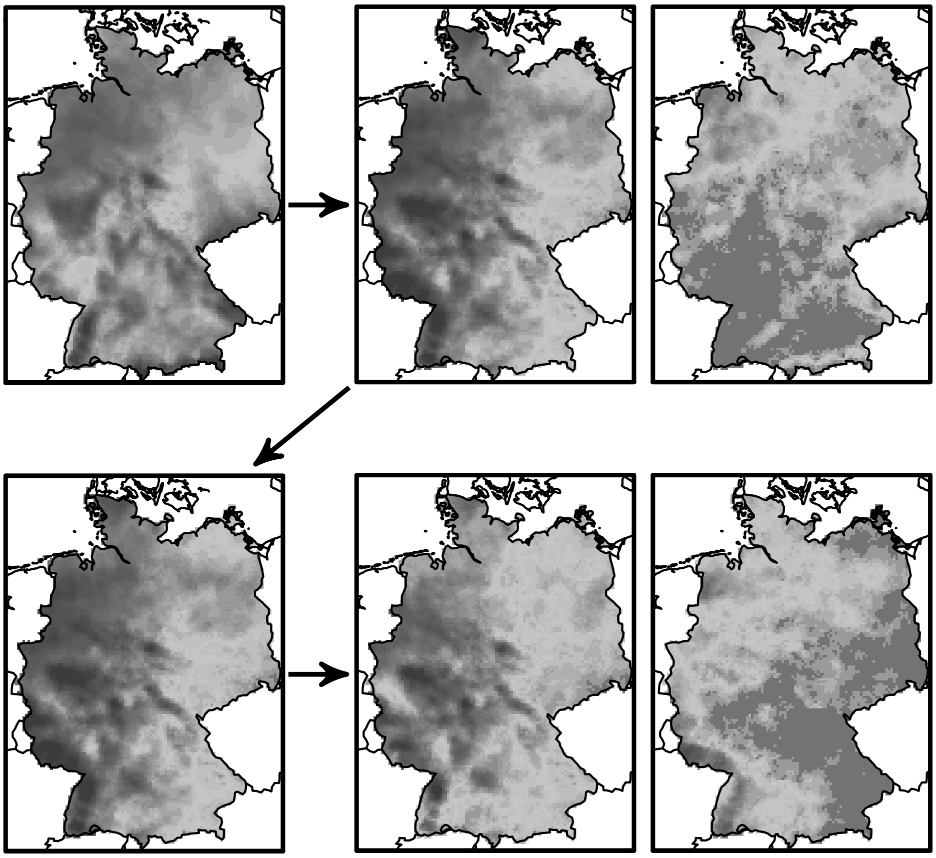
Different ensemble selection methods were compared in their ability to efficiently identify the highest performing subensembles. These methods employ different rationales, such as taking only the best-performing models into account (score-based method) or finding the best-performing subensemble by removing models from the full ensemble (backward elimination method). The backward elimination method was the most efficient and robust technique to select subensembles that are useful for impact assessments and policy recommendations. In many cases, the subensembles selected with the backward elimination method correspond to the best possible combination at a given ensemble size. In contrast, weighted ensemble averaging could not identify them. The backward elimination method is general enough to be applicable to any climate model outputs.

The procedure to identify these subensembles is generally as follows. 1) Compute the rejection rate for all subensemble selected by the backward elimination method. As a rule of thumb, discard subensembles that contain less than half of the models. 2) The subensemble exhibiting the lowest $\overline{R}$ can be directly used for policy recommendations. If the $\overline{R}$ value for subensembles substantially decreases when single models are removed from the full ensemble, then these single models can be discarded for impact assessment studies.

This study has shown that appropriately selected subensembles are able to extract more reliable information from a given multi-model ensemble than weighted ensemble averaging or the full ensemble. Subensemble selection has advantages for impact assessments over weighted averaging schemes or Bayesian frameworks because these typically evaluate surrogates of the RCMs. The performance of the selected subensembles in hydrologic impact studies has yet to be investigated. Moreover, further research should incorporate the evaluation of temporal trends into the presented selection methods.

Chapter 3

Stochastic Temporal Disaggregation of Monthly Precipitation for Regional Gridded Data Sets



Aknowledgments:
This study was carried out within the context of the WESS project and was kindly supported by the Helmholtz Impulse and Networking Fund through Helmholtz Interdisciplinary School for Environmental Research (HIGRADE). The authors also acknowledge the collaboration of the German Weather Service DWD for providing the observational data set employed in this study. We are deeply indebted to M. Cuntz, R. Kumar, O. Rakovec (all Helmholtz Centre for Environmental Research - UFZ, Leipzig, Germany), and A. Bárdossy (Department of Hydrology and Geohydrology, University of Stuttgart, Germany) for critical comments and helpful discussion. We would also like to thank the editor and three anonymous reviewers for constructive comments that helped to further improve the manuscript.

published as:
Stephan Thober, Juliane Mai, Matthias Zink, and Luis Samaniego, "Stochastic Temporal Disaggregation of Monthly Precipitation for Regional Gridded Data Sets", Water Resources Research, 50(11): 8714–8735, 2014, ISSN 1944-7973, doi:10.1002/2014WR015930

3.1 Abstract

Weather generators are used for spatio-temporal downscaling of climate model outputs (e.g., precipitation and temperature) to investigate the impact of climate change on the hydrological cycle. In this study, a multiplicative random cascade model is proposed for the stochastic temporal disaggregation of monthly to daily precipitation fields, which is designed to be applicable to grids of any spatial resolution and extent. The proposed method uses stationary distribution functions that describe the partitioning of precipitation throughout multiple temporal scales (e.g., weekly and bi-weekly scale). Moreover, it explicitly considers the intensity and spatial covariance of precipitation in the disaggregation procedure, but requires no assumption about the temporal relationship and spatial isotropy of precipitation fields. A split sampling test is conducted on a high-resolution (i.e., $4 \times 4\,\mathrm{km}^2$ grid) daily precipitation data set over Germany ($\approx 357\,000\,\mathrm{km}^2$) to assess the performance of the proposed method during future periods. The proposed method has proven to consistently reproduce distinctive location dependent precipitation distribution functions with biases less than 5% during both a calibration and evaluation period. Furthermore, extreme precipitation amounts and the spatial and temporal covariance of the generated fields are comparable to those of the observations. Consequently, the proposed temporal disaggregation approach satisfies the minimum conditions for a precipitation generator aiming at the assessment of hydrological response to climate change at regional and continental scales or for generating seamless predictions of hydrological variables.

3.2 Introduction

Distributed hydrological models are used for evaluating, reconstructing, and projecting states and fluxes representing the hydrological cycle. Typical applications are the quantification of floods (Blöschl et al., 2008; Kumar et al., 2010) and droughts (Sheffield et al., 2004; Vidal et al., 2010; Samaniego et al., 2013). They are commonly used as impact models at high spatio-temporal resolutions (e.g., daily and sub-daily time steps at a few kilometer grid) to support water resource management and planning, for climate change impact studies, or for routine hydrological forecasting (Schaake et al., 2007a). In those cases, meteorological forcings for hydrologic models such as precipitation and temperature are often obtained from outputs of climate model (CM) projections (Booij, 2005; Smith et al., 2014) or Numerical Weather Prediction models (NWPs) (Bougeault et al., 2010; Kirtman et al., 2014). Such outputs are commonly provided at spatio-temporal resolutions coarser than those required by impact models (e.g., daily or monthly values at a 1° grid) (Taylor K. et al., 2012; Kirtman et al., 2014). Different spatio-temporal downscaling schemes are available to bridge the gap between the low resolution of climate and the high one of impact models (Maraun et al., 2010).

Coarse global climate model variables can be dynamically downscaled using regional climate models (Maraun et al., 2010). This approach considers the physics of atmospheric processes but often fails to reproduce extreme precipitation amounts that are important for hydrologic modeling (Thober and Samaniego, 2014). Additionally, dynamical downscaling is computationally very expensive and the obtained spatial resolution is still too coarse for impact modeling (D'Onofrio D. et al., 2014).

A computationally efficient alternative to dynamical downscaling is statistical downscaling (Maraun et al., 2010). This approach uses statistical relationships between coarse-scale predictors and fine-scale predictands. One important feature of statistical downscaling is the addition of local-scale variability that cannot be explained by the coarse-scale predictor (von Storch, 1999; Maraun, 2013). Most of these statistical approaches generate one single set of predictands given a specific set of predictors. If multiple realizations of predictands are required given the same input, as necessary in hydrological ensemble prediction applications, then stochastic frameworks are frequently used. These frameworks are often called weather generators (WGs) because they can generate ensembles of high-resolution meteorological fields (e.g., precipitation) which reproduce observed statistical characteristics and are based on the corresponding low-resolution fields.

A substantial amount of research has been conducted for the spatial downscaling of meteorological variables such as precipitation. This study is aiming at the temporal disaggregation of these variables, which is equally important for two reasons. First, temporal disaggregation of monthly CM outputs to daily values offers the possibility to obtain realistic extremes, which is often not achieved by daily CM outputs (Thober and Samaniego, 2014), and simultaneously preserves the low-resolution characteristics given by CMs. Second, multimodel ensemble (MME) approaches have been promoted in the past years to gain a better understanding of the uncertainty associated with modeling climate change (Taylor K. et al., 2012). Because of the increased storage demand related with MMEs, only monthly variables are made available to practitioners in most cases (e.g., Kirtman et al. (2014)), although the intrinsic model time stepping of CMs and NWPs is much smaller. In those cases, practitioners need to apply a temporal disaggregation in order to use these data sets for their applications.

The stochastic temporal disaggregation of variables such as precipitation can also be considered as a WG method. Several kinds of WGs have been proposed in the literature and extensive reviews of the proposed techniques can be found in Maraun et al. (2010) and Haberlandt et al. (2011). Most of these methods focus on precipitation because it exhibits a highly skewed distribution function and a quite complex spatial covariance structure as compared to those of other variables such as temperature. WGs have been used for generating historic precipitation ensembles when calibrated against observations (Clark and Slater, 2006). Existing methods such as Clark and Slater (2006); Wilks (2009); Bárdossy and Pegram (2009); Burton et al. (2010); Paschalis et al. (2013) have also been used for spatial downscaling of precipitation if they are conditioned on coarse-scale predictors (such as CM outputs) (Wilks, 2010, 2012). These methods, however, have not been employed for the temporal disaggregation of precipitation over larger domains. The main reason for this development stems from the fact that the temporal scaling characteristics of precipitation fields are more complex than the spatial ones (Lovejoy and Schertzer, 2010). This, in turn, implies that the temporal disaggregation of precipitation becomes more difficult than its spatial downscaling. Copula techniques (Bárdossy and Pegram, 2009), for example, are not feasible for temporal disaggregation due to the weak temporal stochastic dependency of precipitation fields at various spatial scales.

To ensure the numerical stability during the spatial sampling, existing WGs often assume isotropic precipitation fields, which implies that the spatial covariance between two grid cells only depends on the distance between these two (Paschalis et al., 2013). This assumption, however, might not hold because of orographic effects, which often lead to an inconsistent spatial covariance matrix (Ebtehaj and Foufoula-Georgiou, 2010). If the assumption of isotropic fields is relaxed, the computability of the Cholesky factors of the cross-covariance matrix becomes a necessary condition for the spatial sampling algorithm (Wilks, 1998, 2009). The disadvantage of this practice appears on large grids (e.g., those with more than 20,000 grid cells) for which the Cholesky factors are not guaranteed to be computable. Alternative methods like copula approaches (Bárdossy and Pegram, 2009) are also unfeasible due to the increased sample size required to estimate a copula for each grid cell as compared to the one required to estimate a covariance.

WGs specifically developed for the temporal disaggregation of precipitation, in general, use multiplicative random cascades and focus mainly on sub-daily time scales (Olsson, 1998; Güntner et al., 2001; Deidda et al., 2006; Paschalis et al., 2012). Such models exploit the scale-independent and multifractal properties of precipitation at these particular temporal resolutions (Deidda et al., 2006; Lovejoy and Schertzer, 2010). No much progress has been achieved with respect to the temporal disaggregation of monthly to daily precipitation at a given spatial resolution. For this case, only deterministic methods are available, which essentially resample observations to generate local-scale variability (Day, 1985). These resampling schemes (e.g., the Ensemble Streamflow Prediction approach) are applicable to short-term seasonal forecasting (Wood and Lettenmaier, 2008) but not for climate change studies because they rely on the assumption that past events are representative for future ones.

The statistical relationships required by WGs are always estimated from past observations regardless of the intended application of the algorithm. For climate impact assessment, however, the implicit stationarity of the statistical relationships employed by a given WGs needs to be verified (Charles et al., 1999). The stationarity test, however, is not a standard practice and seldom found in the literature (Hundecha and Bárdossy, 2008).

Considering the current state-of-the-art of available WGs techniques with respect to temporal disaggregation, spatial sampling, and transferability to future periods, three research questions constitute the main goal of this study. 1) How to generate an ensemble of daily precipitation fields given monthly estimates over a large spatial domain that preserves the observed statistical properties? 2) How reliable is such a disaggregation approach for downscaling CM or NWP outputs during future periods when the temporal disaggregation algorithm is calibrated during past periods? And 3) How to guarantee the numerical stability of a spatial sampling approach for large-scale precipitation data sets without assuming isotropic precipitation fields?

To address these questions, a multiplicative random cascade approach frequently used for the generation of rainfall fields at sub-daily scales (Olsson, 1998; Güntner et al., 2001; Paschalis et al., 2012) is enhanced for devising a computationally efficient algorithm for seamless stochastic disaggregation of monthly to daily precipitation fields. It is then tested whether the disaggregated daily fields resemble the spatial covariance and statistical moments of the observations. We hypothesize that the distribution function for the partitioning of the precipitation amount between low and high temporal resolutions at a given location is quasi-stationary. In contrast, assuming stationarity of the moments of precipitation is a much stronger presumption, which is typically made by other existing techniques (Wilks, 2009) although it does not hold during future periods due to climate change (Kirtman et al., 2013). The quasi-stationarity of the partitioning, on the contrary, assumes that the physical process of precipitation generation is similar along time. For this reason, a method exploiting this property for temporal disaggregation of precipitation fields is considered advantageous with respect to those methods assuming any form of stationarity of moments.

A modified sequential sampling algorithm based on the theoretical work by Dimitrakopoulos and Luo (2004) is proposed to generate precipitation fields that resemble the spatial cross-covariance of observed fields over domains of any size. This novel technique consequently assures the numerical stability of the spatial sampling and exhibits a high computational efficiency without requiring the assumption of isotropic precipitation fields and the explicit Cholesky factorization of the cross-covariance matrix, which becomes numerically intractable for large domains.

The proposed method is evaluated with observed rainfall data over Germany at a high spatial resolution of 4×4 km^2 resulting in approximately 23 000 grid cells. A split sampling test is employed to evaluate the effectiveness of the proposed algorithm during future periods. To achieve this goal, all statistics required by the proposed WG are estimated during the calibration period from 1950 to 1990 and then applied during the period from 1991 to 2010 for evaluation.

3.3 Method for Seamless Stochastic Temporal Disaggregation

In this study, a method is proposed for the temporal disaggregation of monthly precipitation to daily values that is consistent throughout multiple temporal resolutions (e.g., 1, 2,. . ., 32 day; hence seamless) and applicable to large domains (e.g., larger than 10^5 km^2). The proposed method is based on a multiplicative cascade approach which is frequently used on sub-daily scales (Olsson, 1998; Güntner et al., 2001; Paschalis et al., 2012) and is described by

$$p_i^k(t) = \begin{cases} w_i^K(T)\, p_i^K(T) & \text{if}\, t = 2T, \\ (1 - w_i^K(T)) p_i^K(T) & \text{if}\, t = 2T - 1 \end{cases} \tag{3.1}$$
$$\text{for}\, i = 1, \dots, N;\; t = 1, \dots, S;\; K = 2^R, \dots, 2^2, 2^1;\; k = \frac{K}{2},$$

where $p_i^K(T)$ denotes a precipitation value at temporal resolution K, time point T, and cell i; N is the number of cells; R is the number of temporal resolutions; S is the number of time steps at resolution k; and w_i^K is a weight between 0 and 1. A diagram of a downscaling step from resolution K to k is shown in Figure 3.1. As can be seen in this Figure, the number of time steps is doubled from resolution K to k. Furthermore, this approach guarantees a mass conservation of the precipitation amount between different temporal resolutions because it directly follows from equation 3.1 that

$$p_i^k(2T - 1) + p_i^k(2T) = p_i^K(T). \tag{3.2}$$

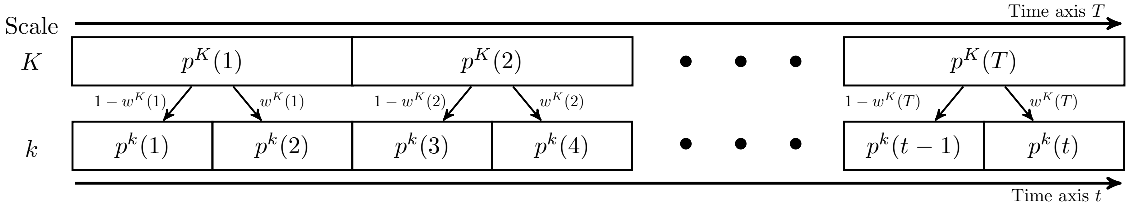


Figure 3.1: Diagram for disaggregation of a precipitation time series at scale K to a precipitation time series at scale k.

The purpose of the weight w_i^K (equation 3.1) is to quantify the partitioning of a given precipitation amount at a scale K (e.g., one monthly value) into two parts at scale k (e.g., two biweekly values). The weight w_i^K for each cell is randomly sampled from a predefined distribution function taking the spatial dependency of the weights into account as given by

$$w_i^K(t) \sim \mathcal{F}_{ic}^K, \tag{3.3a}$$

$$cov\left[g(w_i^K), g(w_j^K)\right] = \mathcal{B}_{ij}^K, \tag{3.3b}$$

$$\text{for } i, j = 1, \ldots, N; k = 2^R, \ldots, 2^2, 2^1$$

where $\mathcal{F}_{ic}^K$ is the cumulative distribution function (cdf) of the weights for the temporal resolution K, cell i, and intensity class c. The distribution functions $\mathcal{F}_{ic}^K$ describe the disaggregation relationships between the precipitation values at consecutive scales K and k. Intensity classes are required because the cdfs $\mathcal{F}_{ic}^K$ depend on the precipitation amount p_i^K that is disaggregated (i.e., distinctive cdfs $\mathcal{F}_{ic}^K$ can be found for low and high precipitation amounts (Güntner et al., 2001)). $\mathcal{B}_{ij}^K$ is the cross-covariance matrix of standard normal variates that are deduced from the weights through a transformation function g. g is a quantile matching function and will be further described in equation 3.6. $\mathcal{B}_{ij}^K$ ensures a preservation of the spatial precipitation covariance when the sampling of the weights is performed.

Following this method, the disaggregation of precipitation is performed from a coarser temporal resolution to a finer one. For example, equation 3.1 is first applied for monthly values (e.g., $k = 16$ and $K = 32$) and then repeatedly applied for decreasing k until daily values ($k = 1$) are obtained.

The whole procedure can be separated into two parts: 1) preprocessing and estimation of $\mathcal{F}_{ic}^K$ and $\mathcal{B}_{ij}^K$ from the given data, and 2) subsequent generation of new realizations employing the statistical properties of $\mathcal{F}_{ic}^K$ and $\mathcal{B}_{ij}^K$. To simplify the notation, indices like K, i, and c for temporal resolution, location, and intensity class, respectively, are going to be suppressed in the following if they are not explicitly required.

3.3.1 Preprocessing and Estimation of $\mathcal{F}$ and $\mathcal{B}$

Different preprocessing steps need to be carried out to set up the generation framework. First, the temporal relationships and the spatial structure of the observed field have to be estimated. A flowchart of these steps is shown in Figure 3.2 and the steps are as follows:

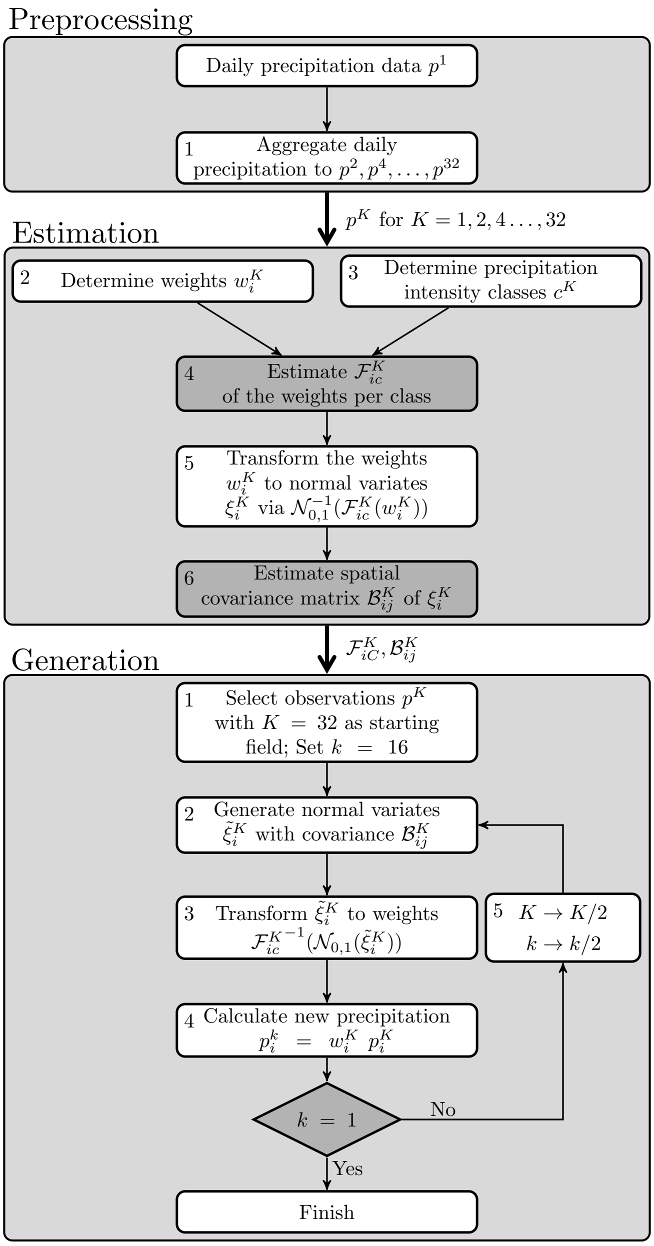

Preprocessing
Daily precipitation data p^1
1 Aggregate daily precipitation to $p^2, p^4, \ldots, p^{32}$
p^K for $K = 1, 2, 4 \ldots, 32$
Estimation
2 Determine weights w_i^K
3 Determine precipitation intensity classes c^K
4 Estimate $\mathcal{F}_{ic}^K$ of the weights per class
5 Transform the weights w_i^K to normal variates ξ_i^K via $\mathcal{N}_{0,1}^{-1}(\mathcal{F}_{ic}^K(w_i^K))$
6 Estimate spatial covariance matrix $\mathcal{B}_{ij}^K$ of ξ_i^K
$\mathcal{F}_{iC}^K, \mathcal{B}_{ij}^K$
Generation
1 Select observations p^K with $K = 32$ as starting field; Set $k = 16$
2 Generate normal variates $\tilde{\xi}_i^K$ with covariance $\mathcal{B}_{ij}^K$
3 Transform $\tilde{\xi}_i^K$ to weights $\mathcal{F}_{ic}^{K^{-1}}(\mathcal{N}_{0,1}(\tilde{\xi}_i^K))$
5 $K \to K/2$ $k \to k/2$
4 Calculate new precipitation $p_i^k = w_i^K \, p_i^K$
$k = 1$
No
Yes
Finish

Figure 3.2: Preprocessing required for the proposed method as well as algorithmic steps required for the estimation and the generation part of the proposed method. The estimation steps are performed for every temporal scale $K = 1, 2, 4, \ldots, 32$.

1.) Sum the observed daily values ($k = 1$) to coarser temporal resolutions $k = 2^1, 2^2, \ldots, 2^R$, for example up to the monthly scale ($R = 5$).

2.) Determine the weights w^K between consecutive temporal resolutions k and K with

$$w^K(t) = \begin{cases} \frac{p^k(t)}{p^K(T)} & \text{if}\, t = 2T, \\ 1 - w^K(2T) & \text{if}\, t = 2T - 1 \end{cases} \tag{3.4}$$
$$\text{for}\, t = 1, \ldots, S;\; K = 2^R, \ldots, 2^2, 2^1;\; k = \frac{K}{2},$$

which can be directly deduced from equation 3.1.

3.) Determine precipitation intensity classes c for a temporal resolution K over the whole spatial domain such that each class has approximately the same number of data points over all cells. The goal of this separation scheme is to capture the distinctive partitioning characteristics of low, medium, and high precipitation intensities. To achieve this, each class is set up to contain 5% of the available data at a given temporal scale k. In addition, each class is constrained to have at least 150 data points to guarantee a reliable estimation of $\mathcal{F}$. This is an important condition for coarser temporal resolutions (e.g., 16 and 32 day precipitation) where only few time steps are available. A detailed description regarding the estimation of the intensity classes can be found in the supplementary material.

4.) Derive the intensity class c of the coarse scale precipitation value $p^K(T)$ and assign it to the weight $w^K(T)$ determined in equation 3.4. Calculate the empirical cdfs $\mathcal{F}_{ic}^K$ of the weights w^K for each cell i, temporal scale K, and precipitation intensity class c using

$$\mathcal{F}_{ic}^K(W) = \mathbf{P}\left(w_i^K \leq W | w_i^K \in c\right), \tag{3.5}$$

where $\mathbf{P}$ denotes the probability of obtaining a weight equal or less than a number W, which is between zero and one.

5.) Transform the weights w^K to their corresponding standard normal variates ξ^K with

$$\xi_i^K = g\left(w_i^K\right) = \mathcal{N}_{0,1}^{-1}\left(\mathcal{F}_{i,c}^K\left(w_i^K\right)\right), \tag{3.6}$$

where $\mathcal{N}_{0,1}$ denotes the normal distribution with mean 0 and variance 1. ξ_i^K is an auxiliary variable that allows to estimate the spatial structure for the partitioning independent of intensity classes c. This relationship g is the transformation function employed in equation 3.3b.

6.) Estimate the cross-covariance matrix $\mathcal{B}^K$ between different locations at temporal resolution K using

$$\mathcal{B}^K = \text{cov}(\xi^K, \xi^K), \tag{3.7}$$

where ξ^K are the fields obtained in step 5).

The major results of the pre-processing part are the cdfs $\mathcal{F}$ for each cell and intensity class and the cross-covariance matrices $\mathcal{B}$ both at every temporal resolution. The cdfs $\mathcal{F}$ are estimated for every location and intensity class c to investigate the maximum efficiency of the proposed disaggregation method. Empirical estimates are used to calculate the cdfs $\mathcal{F}$ cell-wise for the entire domain, because they can be directly computed. To achieve a parsimonious parametrization, theoretical cdfs with regionalized parameters are required to circumvent an optimization problem with too many degrees of freedom. However, the empirical functions estimated in this study are quite similar over space and a few number of clusters could be found to allow a parsimonious parametric representation (see supplementary material, Figure C.2). Such a regionalization would deteriorate the performance at the grid cell level, but reduce the model complexity. The presented study provides a benchmark for these approaches to be investigated in follow-up studies.

3.3.2 Generation of Precipitation Realizations

The procedure for generating precipitation realizations employs the estimated cross-covariance matrices $\mathcal{B}$ and the cdfs $\mathcal{F}$. Figure 3.2 illustrates the single steps used in the proposed method. In detail, these steps are the following:

1.) Fix a temporal resolution k, such that a precipitation field p^K at scale $K = 2k$ is available. For example, one starts with a precipitation field at the monthly scale ($K = 32$) and targets a temporal resolution at the biweekly scale ($k = 16$).

2.) Generate normal distributed values $\tilde{\xi}^K$ with the given covariance matrix $\mathcal{B}^K$. A new sequential Gaussian sampling strategy, called anchor sampling, is applied to assure numerical stability during the generation of $\tilde{\xi}^K$. This new approach conditions the value generated at a specific cell on two kinds of cells. First, a set of neighboring cells which preserves the local spatial structure that has the strongest impact on the generated cell. Second, a set of anchor cells (typically much smaller than the set of neighboring cells) that are equally distributed over the whole domain to represent the global spatial structure correctly. The method is described in detail in Appendix A.

3.) Transform the generated Gaussian random field $\tilde{\xi}^K$ to their corresponding weights w^K as

$$w_i^K = \mathcal{F}_{ic}^{K^{-1}} \left(\mathcal{N}_{0,1} \left(\tilde{\xi}_i^K \right) \right) . \tag{3.8}$$

This equation is the inverse to equation 3.6 and implies that the transformation depends on the cell i, the temporal resolution K, and the intensity class c. The current intensity class is derived from the given precipitation field p^K (step 1.).

4.) Multiply the weights obtained in step 3.) with the available precipitation field p^K to derive the target precipitation field p^k using equation 3.1. Exploiting the property of mass conservation (equation 3.2), it is sufficient to apply steps 3.) and 4.) only for the even time steps and derive the odd ones as remainder.

5.) Repeat step 1.)-4.) for a finer temporal resolutions (i.e., $K \to K/2$ and $k \to k/2$) until the target resolution is daily values (i.e., $k = 1$).

It is worth mentioning that the weights w are consistently generated in space using the cross-covariance matrices $\mathcal{B}$, but independently in time. The cascade approach used here, implicitly induces the temporal correlation. For example, if the precipitation at a coarser scale is dry then the corresponding finer resolution values are also dry.

3.4 Data Sets and Study Domain

The basic requirement for the proposed method is a given gridded data set with a fixed spatio-temporal resolution. In this study, the spatio-temporal resolution of the data set is daily values covering the period from 1950 to 2010 at a $4 \times 4\,\mathrm{km}^2$ spatial grid, resulting in more than 23 000 cells to cover the domain of Germany. This data set was derived by interpolation of rain gauge data ($\geq$ 5000 stations) from the German Weather Service (DWD) employing external drift kriging (Samaniego et al., 2013). This data set is only used as a test case for the presented method. In general, the gridded data set could be derived from various data sources such as rain gauges, weather radar, satellite based remote-sensed precipitation, or climate or NWP model outputs.

The area of Germany is characterized by three main topographical regions: the North German Lowlands, small mountain ranges in Central and Southern Germany ($\leq$ 1000 m elevation), and high mountains of the Alps at the Southern border ($\leq$ 3000 m elevation). The complex topography of Germany induces a wide spread in long-term annual precipitation ranging from less than 600 mm in the Lowlands up to more than 2000 mm in the Alpine regions (Figure 3.3). Six contrasting locations have been selected to show more detailed results. These are Schleswig, Dortmund, and Berlin, which are located in the North German Lowlands (Figure 3.3: locations (1), (2), and (4), respectively). Berlin is exhibiting a more continental climate with less precipitation amounts compared to Dortmund and Schleswig both dominated by marine climate. Further locations are mountainous sites in Central and Southern Germany, i.e., the Brocken and the Feldberg (Figure 3.3: locations (3) and (5), respectively). The latter has an elevation of 1500 m and represents an alpine climate. Moreover, Munich is representative of the Prealps region in Southern Germany (Figure 3.3: locations (6)).

Because of the orographic features as well as different hydro-climatic regimes, the area of Germany provides well suitable test conditions for a stochastic precipitation generator.

3.5 Results and Discussion

3.5.1 Estimation of Intensity Classes C and Distributions $\mathcal{F}$

The available gridded data is split into two disjoint sets for the calibration and the evaluation of the proposed algorithm. The first one comprises the period from 1950 to 1990 and the second one from 1991 to 2010. The cdfs of the weights $\mathcal{F}$ and the cross-covariance matrices $\mathcal{B}$ are estimated only during the calibration period (Section 3.3.1).

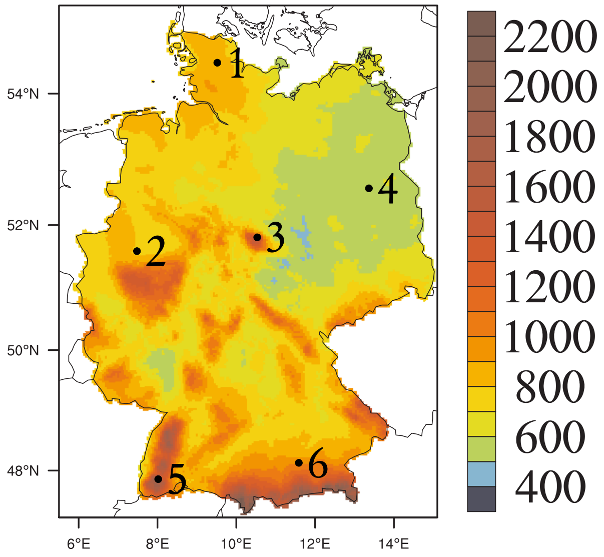


Figure 3.3: Map of long term annual precipitation [mm] across Germany including the positions of six locations representing different precipitation regimes: (1) Schleswig, (2) Dortmund, (3) Brocken, (4) Berlin, (5) Feldberg, and (6) Munich.

These functions depend on the intensity classes c which are common for the whole domain and are estimated beforehand (Table 3.1). The maximum number of classes is reached at the eight day scale with 19 intensity classes. The number of classes decreases for lower temporal resolutions (e.g., daily scale) because of the large spatial variability of precipitation cdfs at these scales and because of the constraint that the intensity classes are constant over space. For higher temporal resolutions (e.g., 32 day scale), the number of classes becomes even smaller because of the fewer number of available time steps and the additional constraint that each class should have at least 150 data points.

As an example, the cdfs $\mathcal{F}$ of the weights to disaggregate two day to daily precipitation for two contrasting locations and for 14 intensity classes are shown in Figure 3.4, one located in lowlands (Figure 3.4a) and the other in highlands (Figure 3.4b). There is a large spread of the cdfs for the different intensity classes because the lower intensity ones (up to 1 mm) have a higher probability of drawing weight equal zero as compared to the high intensity ones (say, larger than 7.4 mm). This is a crucial property for modeling the probability of dry days adequately. Such a distinctive partitioning for low and high precipitation intensities can not be achieved with a single intensity class (gray line in Figure 3.4). Also, two intensity classes as proposed by Güntner et al. (2001) for sub-daily scales would be insufficient (results not shown).

The cdf for the highest intensity class at the highland location exhibits a significant gap with respect to the other cdfs (dark blue line in Figure 3.4b). For example, it attributes around 5% probability to weights equal zero whereas all the others assign at least 10%. This can be seen as a shortcoming of keeping the intensity classes constant over space. A better fit of the cdfs at this location could be achieved by determining the intensity classes individually at each grid cell. This, in turn, would imply an increase of the number of intensity thresholds by a factor of 23 000, which would lead to an

class	2d scale	4d scale	8d scale	16d scale	32d scale
1	0.2	0.4	0.8	5.8	31.3
2	0.5	0.8	2.0	10.5	45.5
3	0.9	1.4	3.3	14.7	58.9
4	1.4	2.2	4.6	18.8	74.4
5	2.0	3.0	6.0	23.1	96.8
6	2.7	3.9	7.5	27.6	158.9
7	3.5	5.0	9.1	32.6	713.0
8	4.6	6.1	10.7	38.2	
9	5.8	7.4	12.5	44.8	
10	7.4	8.8	14.4	53.3	
11	9.5	10.5	16.4	66.2	
12	12.5	12.5	18.7	97.4	
13	18.2	15.0	21.3	467.2	
14	282.5	18.1	24.3		
15		22.8	27.9		
16		31.4	32.4		
17		364.1	38.8		
18			50.1		
19			395.9		

Table 3.1: Upper thresholds [mm] for precipitation intensity classes c estimated for the observed data set for each temporal resolution during the calibration period 1950-1990. The lower threshold of a particular class is the upper threshold of the preceding class. The lower threshold of the first class is zero.

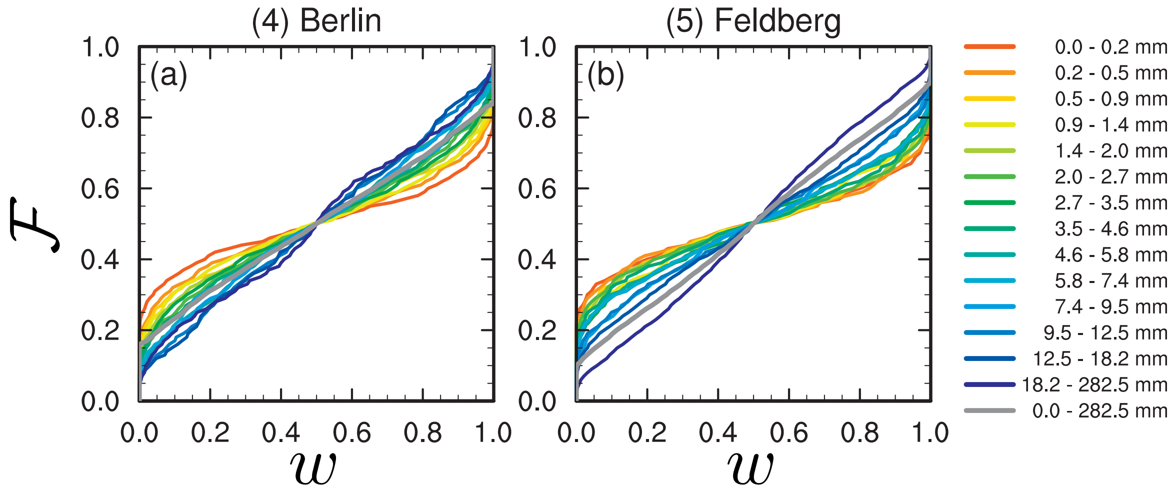


Figure 3.4: Cumulative distribution function (cdf) $\mathcal{F}$ of the weights w between the two-day and daily resolution for (a) location (4) Berlin and (b) location (5) Feldberg estimated during the calibration period of 1950-1990. Colored lines indicate the cdfs for different intensity classes (Table 3.1) ranging from low (red) to high intensities (blue). Gray lines show the cdfs assuming only one intensity class.

overparameterized and impractical method. In this study, spatially constant thresholds are preferred because the cdfs $\mathcal{F}$ are already location dependent.

Empirical evidence supporting the hypothesis related with the stationarity of the cdfs $\mathcal{F}$ is depicted in Figure 3.5 for the six selected locations. In this figure, a quantile-quantile plot of empirical cdfs $\mathcal{F}$ for the disaggregation of two-day to daily precipitation estimated during the calibration and evaluation period is shown. Most of the lines exhibit a close match with the 1:1-line, which implies that the cdfs $\mathcal{F}$ have not changed from the calibration to the evaluation period. In addition, the non-parametric Kolmogorov-Smirnov test under the null hypothesis that empirical cdfs $\mathcal{F}$ during the calibration and evaluation periods are the same cannot be rejected for middle and high intensity classes at any location at a 5% significance level. At the low intensity class (0.9-1.4 mm), the deviations from the 1:1 line are, in general, larger than for the other classes which indicates that a shift in the cdfs $\mathcal{F}$ from the calibration to the evaluation period has occurred. At high altitude locations (3) and (5), these mismatches, however, may be related to the comparatively small sample size during the evaluation period leading to a wrong estimate of the cdf $\mathcal{F}$.

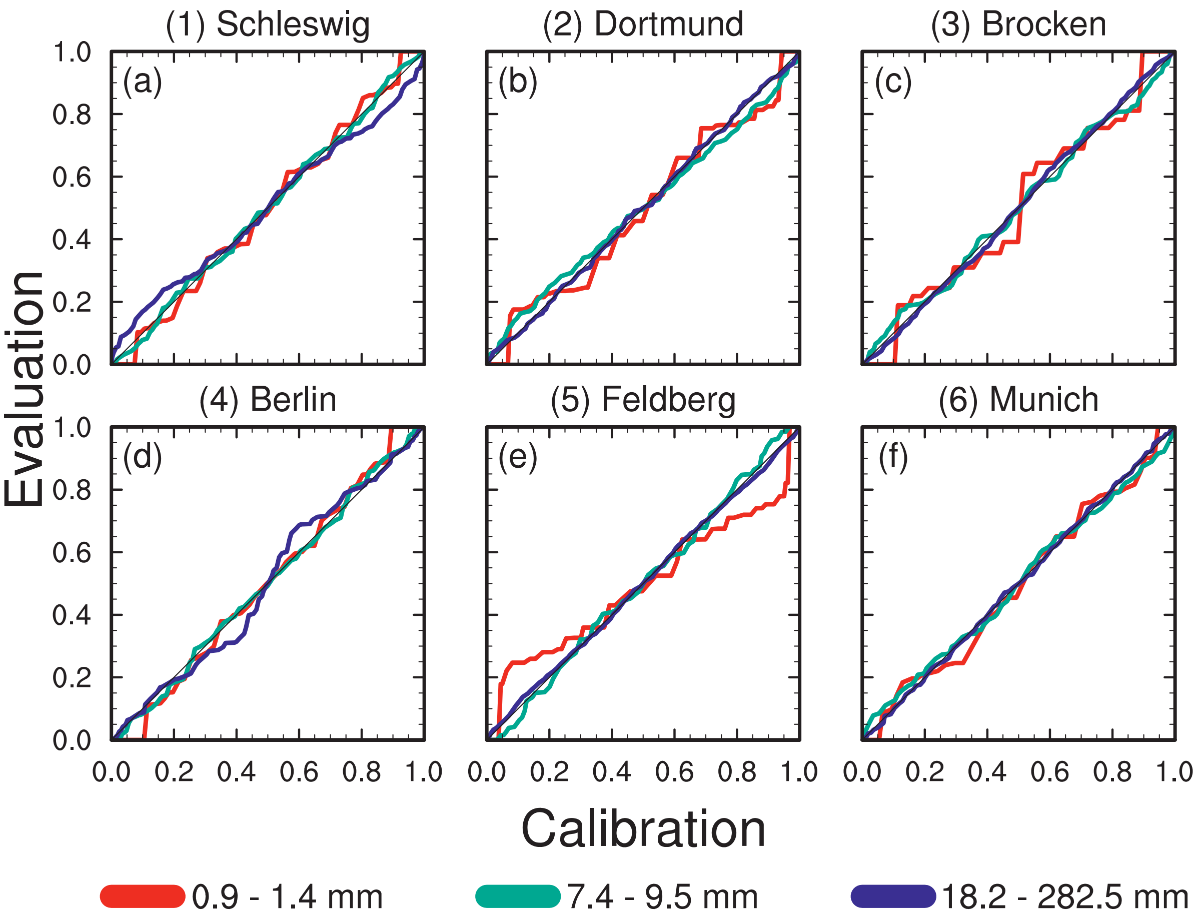


Figure 3.5: Quantile-Quantile plot between the cdfs $\mathcal{F}$ for the partitioning of two-day precipitation estimated during the calibration and evaluation period at six contrasting locations. The cdfs are depicted for three selected intensity classes ranging from low (red lines) to high ones (blue lines).

In general, several intensity classes c are required to capture the distinctive partitioning of low, medium, and high precipitation intensities. Additionally, the assumption of stationary cdfs $\mathcal{F}$ holds quite well in the study domain for middle and high intensities, which are more important for rainfall-runoff modeling as compared to low ones.

3.5.2 Verification of Precipitation Distribution Functions

The proposed method is used to stochastically disaggregate observed monthly precipitation during the calibration and evaluation period based on empirical cdfs $\mathcal{F}$ and cross-covariance matrices $\mathcal{B}$ estimated during the calibration period. One exemplary realization of a disaggregation cascade of a 32 day field is depicted in Figure 3.6. The proposed method conserves the precipitation amount, implying that the sum of the two disaggregated fields (i.e., two right fields) equals the field at the coarser time step (i.e., left field). A spatial clustering of rainfall areas is achieved by the proposed method, which can be attributed to the sequential sampling algorithm described in Appendix A. Larger dry areas (blue cells in Figure 3.6) can be observed when the temporal resolution is increased, in particular from the four to the two day resolution. This is due to the multiplicative cascade model structure of the proposed method because a dry cell at a certain temporal resolution will also be dry at all higher temporal resolutions.

The proposed method is utilized to generate two 100 member ensembles (one for each the calibration and evaluation period). Disaggregated daily precipitation intensities are then compared against the observed values. It is worth mentioning that the observed data set is only used for validation of the proposed method during the evaluation period. In general, the estimated statistics $\mathcal{F}$ and $\mathcal{B}$ can be used for the disaggregation of various monthly data sets like climate or NWP model outputs such as Kirtman et al. (2014).

Different percentiles covering low and high positive precipitation amounts are calculated for the observations and the simulations at each cell. These control points correspond to the deciles $D_1, \ldots, D_9$ and three high percentiles on the upper tail for characterizing extreme precipitation intensities. The selected high percentiles are the 95^{th}, 99^{th}, and 99.9^{th} percentile.

The performance of the method is assessed by quantifying the mean absolute relative bias (MARB: equation B.2 in Appendix B) between the median of the simulated and the observed value over given statistics (e.g., deciles, high percentiles). Additionally, the 95% confidence interval of the simulations is determined at each grid cell for all statistics mentioned above to test the null hypothesis that the simulated and observed values are the same. The null hypothesis is rejected if an observed value is outside of the 95% confidence interval. The rejection rate (RR) is estimated as the ratio between the number of rejected statistics and the total number of tests.

During the calibration period, 92% of the cells exhibit a RR less than 40% (Table 3.2a). Cells exhibiting RR greater than 40% are randomly distributed over Germany (Figure 3.7a). The MARB for the deciles is, in general, less than 5% (Figure 3.7b). This bias is comparable to the measurement error of rain gauges operated by the German Weather Service (MANOB, 2006; Sieck et al., 2007) and thus emphasizes the reliability of the proposed method. Moreover, this bias is quasi-constant in space which means that it can be associated with the noise in the random number sampling, which is intrinsic to every stochastic generation framework.

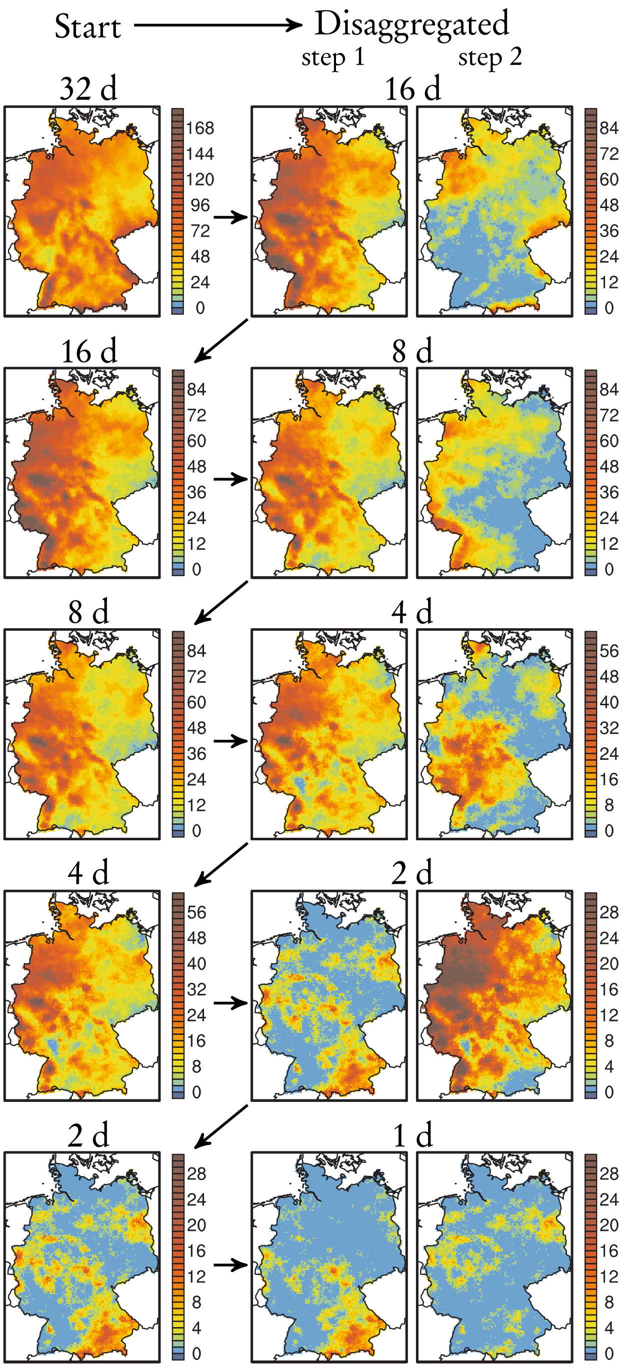


Figure 3.6: One exemplary disaggregation cascade for a single 32 day field. The unit throughout the panels is [mm]. The top left plot is an observed 32 day value. It is depicted in each row how a low resolution value (left plot) is stochastically disaggregated to two higher resolution values (right plots).

The MARB for the high percentiles is significantly larger than those of the deciles during the calibration period (compare Tables 3.2b and 3.2d for the calibration period). It should be noted, never-

Deciles

a)

Rejection rate less equal x%	Percentage of cells calibration	Percentage of cells evaluation
0	28	22
20	59	41
40	92	69
60	99	86
80	100	96

b)

MARB less equal x%	Percentage of cells calibration	Percentage of cells evaluation
1	0	0
2.5	49	27
5	100	77
7.5	100	94
10	100	99

High percentiles

c)

Rejection rate less equal x%	Percentage of cells calibration	Percentage of cells evaluation
0	44	61
40	87	94
80	99	100

d)

MARB less equal x%	Percentage of cells calibration	Percentage of cells evaluation
5	36	28
10	73	70
20	98	98

Table 3.2: Percentage of cells exhibiting a given level of (a) rejection rate (RR) and (b) mean absolute relative bias (MARB) with respect to all deciles. Tables (c) and (d) are the same as (a) and (b), but for high percentiles, which are the 95^{th}, 99^{th}, and 99.9^{th} percentile.

theless, that these biases are dominated by heavy precipitation events, which occur very rarely. Most cells (87%) exhibit a RR less than 40% (Table 3.2c). Hence, the simulated precipitation at higher percentiles is less precise (i.e., exhibit larger MARB) than those obtained at lower ones, but still exhibit a comparable RR. This characteristic is advantageous for a downscaling technique because a low RR of high precipitation amounts is crucial for hydrological applications, such as flood forecasting.

During the evaluation period, the number of cells exhibiting a given level of RR for deciles is less than the corresponding one obtained during the calibration period (Table 3.2a). In fact, the percentage of cells exhibiting a RR less than 40% decreases from 92% during the calibration period to 69% during the evaluation period. Cells exhibiting a high RR are clustering in Central-Northern Germany (Figure 3.7c), which coincide with the region exhibiting the highest MARB (Figure 3.7d).

The shift in the observed precipitation distribution functions between the calibration and evaluation period, however, is substantially larger (Figure 3.8a). For example, Central-Northern Germany exhibits differences up to 17% and, overall, 55% of the cells exhibit a MARB greater than 5%. Considering the generated precipitation ensemble during the evaluation period, the MARB is only in few locations higher than 10% (Figure 3.8b) and, overall, only 23% of the cells exhibit a MARB greater than 5% (Table 3.2b). This highlights that the proposed method is to a large extent able to model the shift seen in observed precipitation cdfs between the calibration and evaluation period. It should be noted, however, that the change in observed precipitation cdfs between the calibration and the evaluation periods might be more related to differences in the observational data set rather than to climate change because there is a significant decrease in density of operated rain gauges in Northern, Central, and Eastern Germany between these two periods.

Figure 3.9 shows a detailed analysis of observed and simulated precipitation cdfs at the six different locations depicted in Figure 3.3. The proposed disaggregation method is able to reproduce the location dependent precipitation intensities very well because the cdfs $\mathcal{F}$ are estimated for every location individually. Regionalization of the precipitation intensity classes could lead to a better estimate at

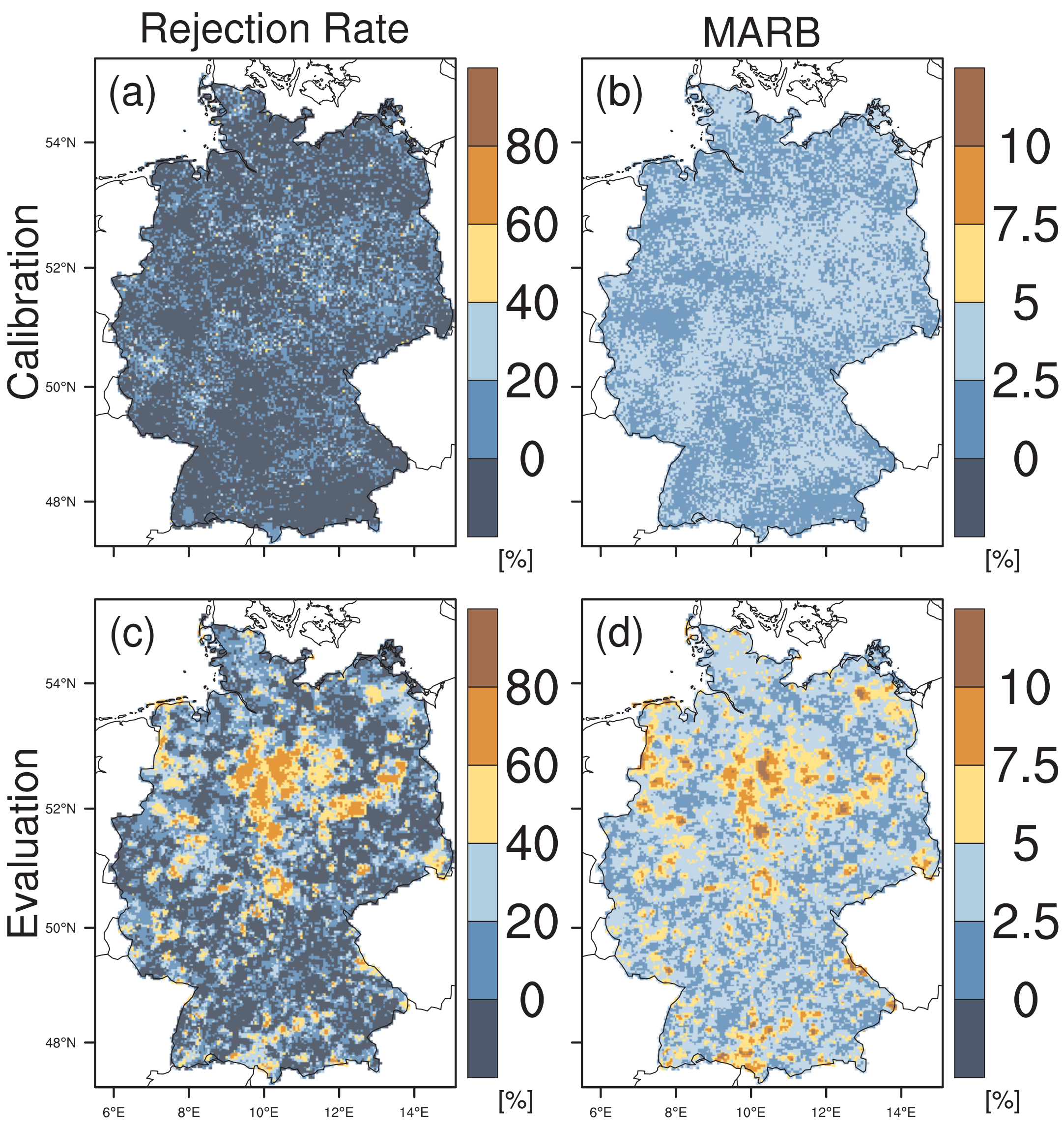


Figure 3.7: Evaluation of deciles of the precipitation distribution function at every grid cell. (a) The rejection rate (RR) and (b) the mean absolute relative bias (MARB) over all deciles $D_1, \ldots, D_9$ during the calibration period. Second row: same as the first row, but for the evaluation period.

mountainous sites, for example for high precipitation intensities at location (5) (Figure 3.9). Other highland locations, however, exhibit an estimate that is as good as that of lowlands, e.g., compare location (2) and (3) in the same Figure.

The 95% confidence interval is very narrow for exceedance probabilities higher than 1% and is widening below this threshold (Figure 3.9). Narrow uncertainty bands are estimated for frequently occurring precipitation intensities such as deciles D1 to D9. A side effect of this characteristic is the rela-

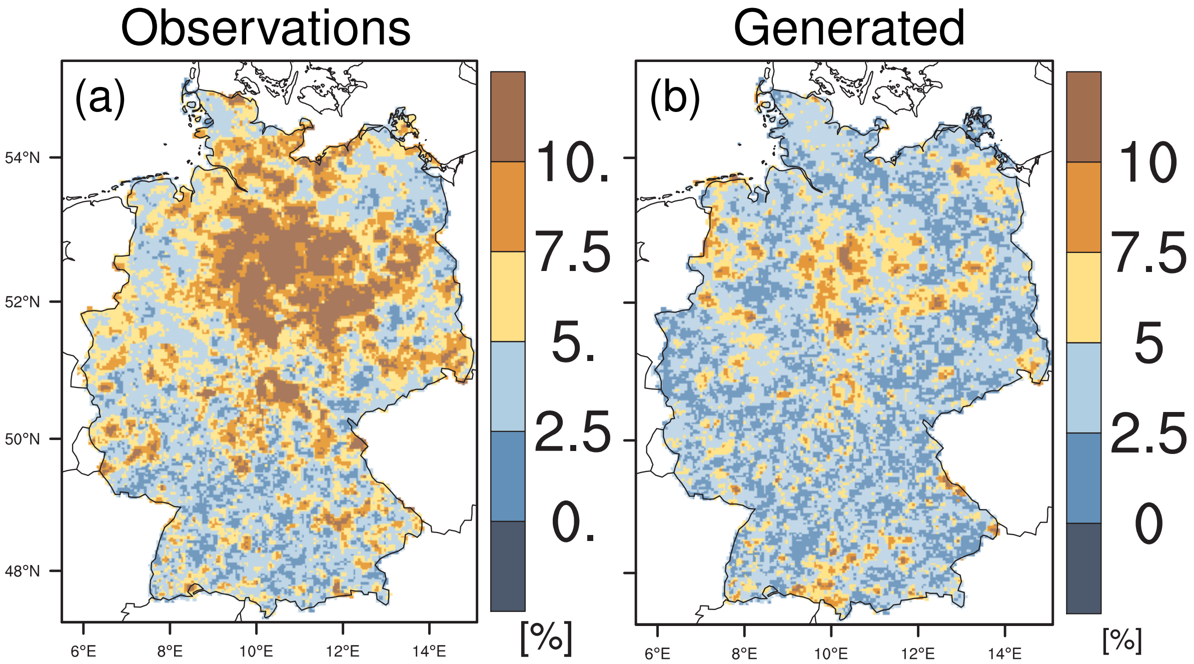


Figure 3.8: Panel (a) shows the differences in observed precipitation distribution functions between calibration and evaluation periods. For comparison, panel (b) shows the difference between the simulation and observation during the evaluation period (same as Figure 3.7d). The differences are calculated as the mean absolute relative bias (MARB) over all deciles $D_1, \ldots, D_9$ of the respective distribution functions.

tively high RR for deciles although the corresponding MARB is less than 10% during the evaluation period (Figures 3.7c and 3.7d).

By design, any stochastic precipitation generator is calibrated on a set of statistics that have to be reproduced as close as possible. Commonly, these statistics are moments of precipitation cdfs (Wilks, 1998; Maraun et al., 2010). For example, if these cdfs change over time, simulated precipitation will diverge from observations. The proposed method is quasi-insensitive to most of these shifts as long as the disaggregation scheme remain invariant over time. A change in the disaggregation scheme could be caused, for example, by shifts in the precipitation regime from frontal to convective ones.

3.5.3 Probability of Dry Days

In addition to the ability of reproducing the cdfs of precipitation greater than zero, a precipitation generator should have the skill in correctly estimating dry precipitation conditions, which are important for drought studies (Sheffield et al., 2004; Vidal et al., 2010; Samaniego et al., 2013). For this purpose, the performance of the proposed method with respect to the probability of dry days ($P0$) is analyzed and depicted in Figure 3.10. A day is defined as dry if the precipitation intensity at that day is less than 0.1 mm.

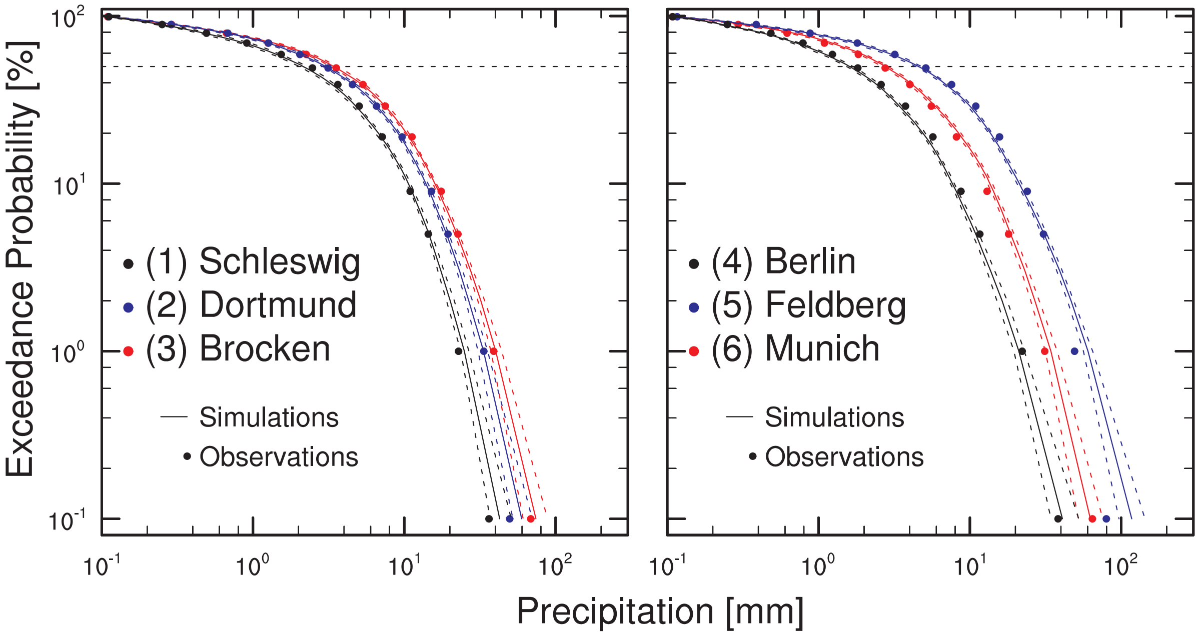


Figure 3.9: Exceedance probabilities of daily precipitation depicted for six selected locations (Figure 3.3) during the evaluation period 1991-2010. Dots mark the observed values at exceedance probabilities of $0.1\%, 1\%, 5\%, 10\%, 20\%, 30\%, \ldots, 100\%$. Solid lines depict the median of simulations and dashed lines the 95% confidence interval. The horizontal dashed line shows the 50% exceedance probability.

In general, the eastern part of Germany and the lowlands in Southern Germany are characterized by relatively higher number of dry days compared to the rest of Germany (Figures 3.10a and 3.10d). This spatial feature is correctly represented by the simulations (Figures 3.10b and 3.10e). The relative biases are generally less than 4% during the calibration period (Figure 3.10c). The performance of the method deteriorates during the evaluation period. The relative biases during this period are between ±8% (Figure 3.10f).

It is worth mentioning that the areas with the strongest over- and underestimation of $P0$ during the evaluation period coincide with those areas exhibiting the highest MARB over all deciles (Figure 3.7d). This is a consequence of the mass conservation of precipitation amounts, which is intrinsic in the proposed method. In other words, if the number of dry days would be overestimated, then positive precipitation amounts would be distributed over fewer days, which leads to an overestimation of precipitation intensity. The other way around is also possible. This interplay links the precipitation cdf with the probability of dry days.

3.5.4 Extreme Precipitation Characteristics

The correct representation of extreme daily precipitation characteristics is crucial for the estimation of both high and low streamflow. The total precipitation amount above the 95^{th} percentile ($r95p$)

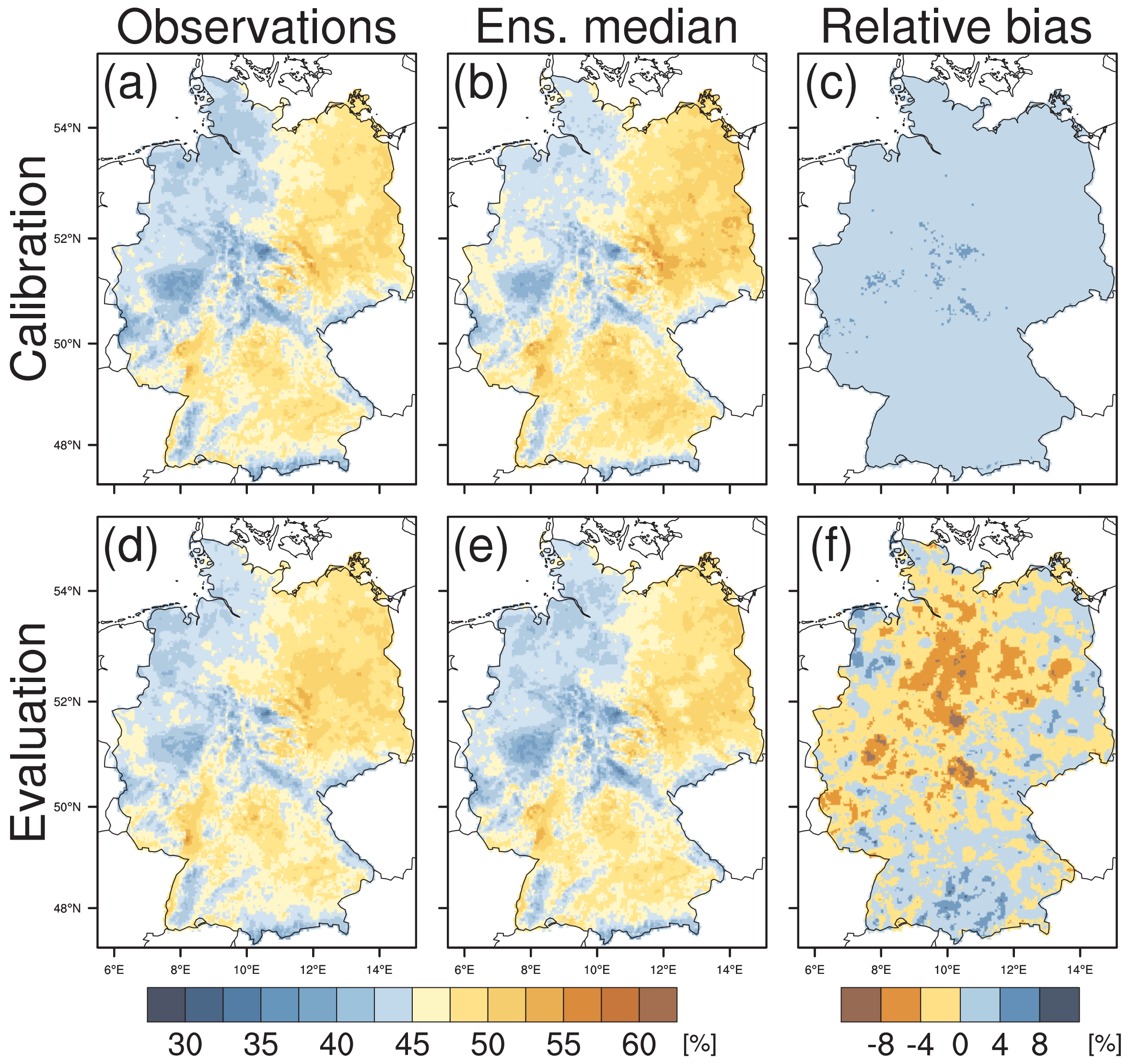


Figure 3.10: Probabilities of dry days for (a,d) the observations and (b,e) the ensemble median of simulations during the calibration (1950-1990) and the evaluation period (1991-2010). The relative bias between the ensemble median of simulations and the observations is shown in panels (c) and (f) during the calibration and evaluation period, respectively.

and the maximum number of consecutive dry time steps (CDT) are selected as evaluation criteria for extreme wet and dry conditions, respectively. $r95p$ quantifies the upper tail of the precipitation cdf whereas CDT indicates the duration of extreme dry spells. Both of these indices have been frequently used for the evaluation of modeled precipitation (Sillmann and Roeckner, 2008; Herrera et al., 2010; Thober and Samaniego, 2014).

In general, the absolute relative bias (ARB: equation B.1 in Appendix B) for CDT is higher than that of $r95p$, reaching values up to 40% and 10%, respectively (Figures 3.11a and 3.11b). This result is a consequence of estimating cdfs $\mathcal{F}$ explicitly for high precipitation amounts whereas the length of dry spells is an emerging characteristic of the proposed method. In particular, regions with high annual precipitation exhibit a relatively higher bias in $r95p$ than other regions but lower bias in CDT (e.g., Black Forest in South-Western Germany). The highest biases in CDT occur in areas exhibiting the shortest dry spells (e.g., Western Germany).

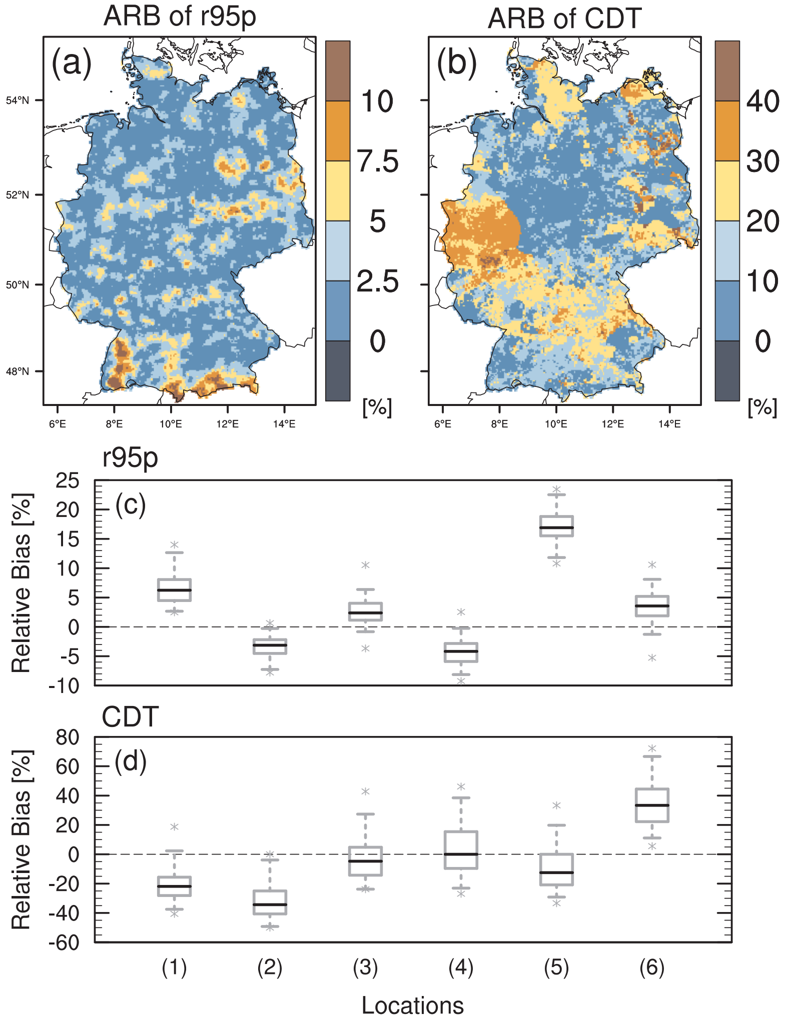


Figure 3.11: Median of absolute relative bias (ARB) between the simulated daily ensemble and the observed daily values for the extreme indices (a) $r95p$ and (b) CDT at every cell during the evaluation period. $r95p$ denotes the sum of rainfall above the 95^{th} percentile and CDT denotes the maximum number of consecutive dry time steps. Boxplots show the ensemble variability of relative bias for (c) $r95p$ and (d) CDT with respect to the observations (dashed line) at six selected locations (Figure 3.3). The median of the ensemble is represented by a bold black line. The box ranges from the 25th to the 75th percentile. The whiskers show the range of the 95% confidence interval and the asterisks mark the minimum and maximum value.

The spatial variability of the ARB of $r95p$ is smoother than that of CDT (compare Figures 3.11a and 3.11b). This feature is related to the sample size used to estimate these indices. While 5% of the wet time steps are considered for calculating $r95p$, only the maximum value of consecutive dry time steps is considered for estimating CDT.

The variability of the ARB for the whole simulated ensemble is shown in detail at six selected locations in Figures 3.11c and 3.11d. The simulated $r95p$ ensemble is not covering the observed value of $r95p$ at locations (1) and (5) (Figure 3.11c), which is caused by a combination of a large bias and lack of spread in the ensemble. These mismatches occur at 11% of the cells. Most of these cells are located in the highlands because the common intensity classes are not necessarily suitable for these. For example, location (5) is one of the highest locations in Germany and the errors obtained for this location represent quite well the maximal error that this algorithm is producing in Germany. This, however, does not imply that most highland locations are exhibiting higher biases. For example location (3) has a median bias of less than 5%.

The ensemble median and spread of the bias for CDT are greater than those obtained for $r95p$ (Figure 3.11d). For example, the ensemble median is exhibiting a large negative bias at locations (1) and (2) of up to -35%. Such a large negative bias tend to occur in regions exhibiting a high ARB (Figure 3.11b). This implies that the proposed method tends to generate dry spells that are too short compared to the observations. This shortcoming is a consequence of the fact that there is no explicit assumption about the temporal dependency of the weights. This method is however able to achieve reasonable estimates at other locations such as (3), (4), and (5). Although the relative bias of CDT are in general greater than those of $r95p$, the percentage of cells where the simulated range does not cover the observed value (e.g., at location (6)) is approximately equal to that obtained for $r95p$ (i.e., 11%). Hence, predicting one index (CDT) with a lower level of precision does not necessarily imply that the rate of mismatches would increase.

3.5.5 Spatial Correlation for different Scales and Seasons

A crucial requirement for a gridded WG is to reproduce the spatial pattern of observed precipitation fields. To test this requirement, the cross-correlation coefficients ρ between simulated time series separated by a given distance are compared against the corresponding ones obtained with the observations.

Six locations, two temporal resolutions, and two seasons are selected for depicting the results of this test (Figure 3.12). As expected, the correlation coefficients ρ decrease with increasing distance as precipitation amounts become more independent of each other. The magnitude of this decrement, however, is location, scale, and season dependent. For example, a sharp decrease of ρ values for distances smaller than 40 km can be observed for mountainous locations (Brocken (3) and Feldberg (5)) whereas a smooth decrease is observed for the other locations lying in lowland areas. Furthermore, the correlation coefficient ρ of daily values decreases more rapidly than that of weekly values (solid blue and red lines, respectively). This stems from the fact that weekly rainfall fields are aggregating daily ones and thus become stronger connected in space.

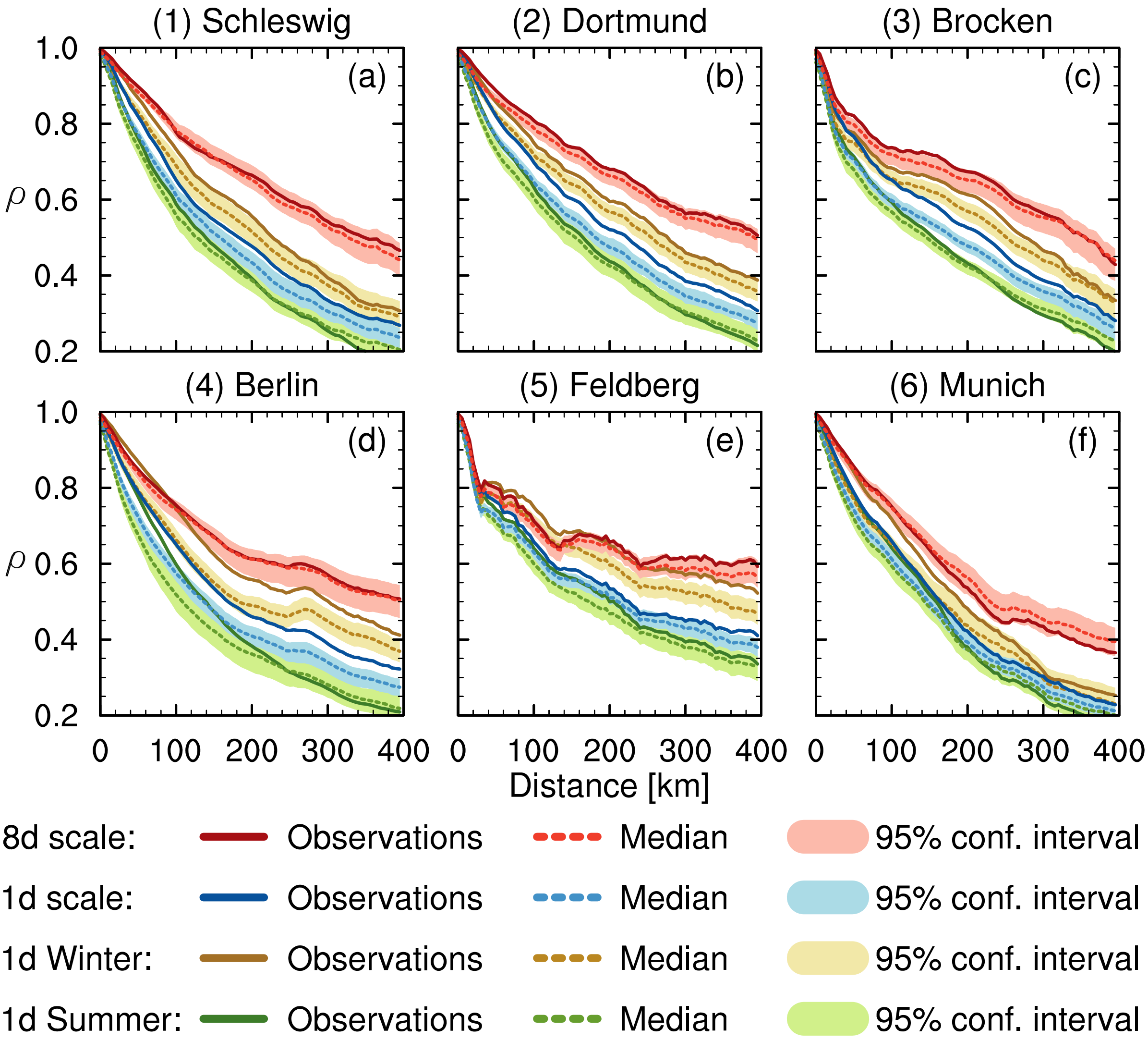


Figure 3.12: Pearson cross-correlation coefficients ρ between specific locations and all other locations averaged over distance in bins of 5 km width during the evaluation period. The Pearson cross correlation ρ is shown for the observations (solid lines) and simulations (dashed lines). Red lines have been calculated for the weekly time scale (8 d). For the daily scale, blue lines have been estimated using the whole data set for all months, brown lines for summer months (MJJASO), and green lines for during winter months (NDJFMA). Additionally, the 95% confidence interval is depicted as bands. The Pearson cross correlation is shown for six selected locations (Figure 3.3).

The cross-correlation ρ estimated for daily values during winter is significantly higher than those estimated using the whole data set and is less than those of weekly values (Figure 3.12: compare brown with blue and red solid lines, respectively), whereas the one estimated during summer is typically the lowest among all considered scales and seasons (green solid line). This is related to the frequency of occurrence of stratiform and convective precipitation. The former distributes precipitation amounts evenly over large areas (i.e., several hundred kilometers). The latter evolves more often during summer when solar radiation is stronger. Convective precipitation only occurs over a very limited spatial extent and thus leads to weaker spatial correlations at larger distances.

For the weekly scale, the observed ρ values are always within the 95% confidence interval of the simulations at all reference locations (red lines in Figure 3.12). For the daily scale, the observed ρ values are however often higher than the simulated ones and lying outside of the 95% confidence interval regardless of the season (blue, brown, and green lines in Figure 3.12). This implies that the simulated precipitation amounts are spatially less correlated than the observations. The intrinsic randomness in the generation procedure causes this lack of performance because it introduces random noise in the simulated fields. This randomness becomes more apparent at the daily time scale because the precipitation amounts exhibit the strongest intermittency at this scale. The proposed method, however, is still able to replicate distinctive location dependent features such as the sharp decrease of the Pearson correlation in the neighborhood of highland locations (Figures 3.12c and 3.12e). It is worth mentioning that this distinctive feature, would not have been achieved if the WG would have assumed an isotropic covariance structure, which is often assumed in literature (Paschalis et al., 2013). Moreover, the method is able to correctly reproduce higher cross-correlations during winter and lower ones during summer, which implies that the method is able to reproduce seasonal changes in the spatial precipitation structure. This is achieved by estimating covariance matrices $\mathcal{B}$ individually during both seasons.

3.5.6 Evaluation at Multiple Temporal Scales

In addition to a realistic representation of observed daily precipitation characteristics, the proposed WG should also preserve these at multi-day resolutions, e.g., at 2, 4, 8, and 16 days. For this purpose, the relative bias between observations and the median of the simulations for different statistics over the whole study domain is analyzed during the evaluation period (Figure 3.13). The investigated statistics are: (a) deciles $D_1, \ldots, D_9$; (b) $r95p$; (c) CDT; and (d) the Pearson cross correlation coefficient ρ.

With regard to the deciles, 95% of the grid cells exhibit a relative bias ranging between -10% and 25% (Figure 3.13a). Among all deciles, the largest absolute values of relative biases are obtained for deciles D_1 and D_2 (results not shown). This is due to the fact that the precipitation amounts at this level are relatively small by value and thus small deviations cause high relative biases. The relative biases for high deciles range between $\pm$5% irrespective of the considered temporal resolutions. These results indicate that the proposed method is able to reproduce precipitation cdfs over different time scales sufficiently well.

In case of the $r95p$, 95% of the grid cells exhibit relative biases ranging from -6% to 10% (Figure 3.13b). Additionally, half of the relative biases are between -2% and 3%, which emphasizes the ability of the proposed method to precisely reproduce high extreme precipitation amounts at multiple time scales. This is due to the explicit consideration of the precipitation intensity classes in the calculation of the cdfs $\mathcal{F}$ (equation 3.5).

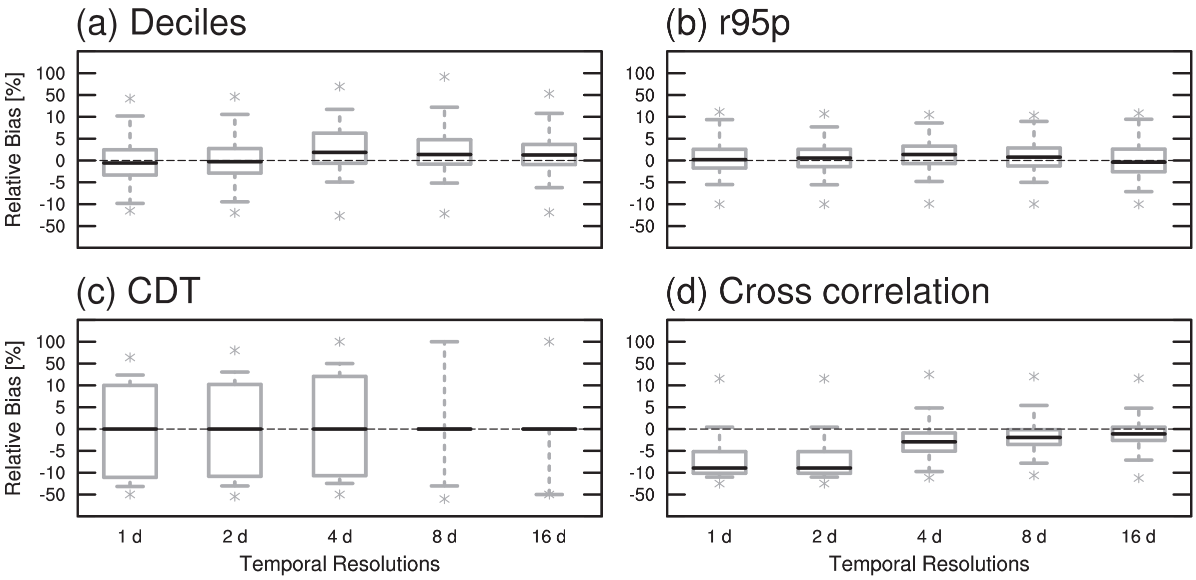


Figure 3.13: Boxplots of relative bias between the median of simulations and observations over all cells during the evaluation period for a specific temporal resolution and statistic. The presented statistics are (a) all deciles, (b) the index $r95p$, (c) the index CDT, and (d) the Pearson cross correlation coefficient ρ. The median of relative bias is marked as black line. The box ranges from the 25^{th} to the 75^{th} percentile. The whiskers show the spread of the 95% confidence interval and the asterisks mark the minimum and maximum value.

The relative bias for CDT ranges from -30% to 50% for 95% of the cells and from -15% to 25% for half of the cells (Figure 3.13c). At low temporal resolutions (e.g., 8 and 16 days), the relative biases are difficult to interpret because of the low CDT values. For example, at the 16 day scale, the median CDT in the observations takes the value of one time step. In this case, a deviation by one unit implies a relative bias of 100%. At high temporal resolutions (e.g., 1, 2, and 4 days), the uncertainty is larger as compared to that of the other statistics because there is no explicit restriction to guarantee a realistic temporal relationship of the weights w. The relative bias for CDT is however symmetrically distributed around its median, which is close to zero.

The range of the relative bias for the Pearson cross-correlation coefficient ρ is between -17% and 5% for 95% of the cell pairs (Figure 3.13d). The median of the relative bias is negative for every temporal resolution, which implies that the proposed method systematically generates fields that are spatially less correlated than the observed ones. In particular, half of the relative biases are between -10% and -5% for high temporal resolutions (e.g., 1 and 2 days). The magnitude of underestimation decreases for lower temporal resolutions (e.g., 4, 8, and 16 days), where half of the relative biases are between -5% and 0%, which implies that the underestimation is negligible at these temporal resolutions.

3.5.7 Temporal Correlation

Since the weights w are sampled independently in time, it has to be tested whether the simulated time series preserve the observed temporal structure of precipitation. As an example, the Pearson auto correlation coefficients γ of daily precipitation for lag times up to ten days are analyzed at six selected locations during the evaluation period (Figure 3.14).

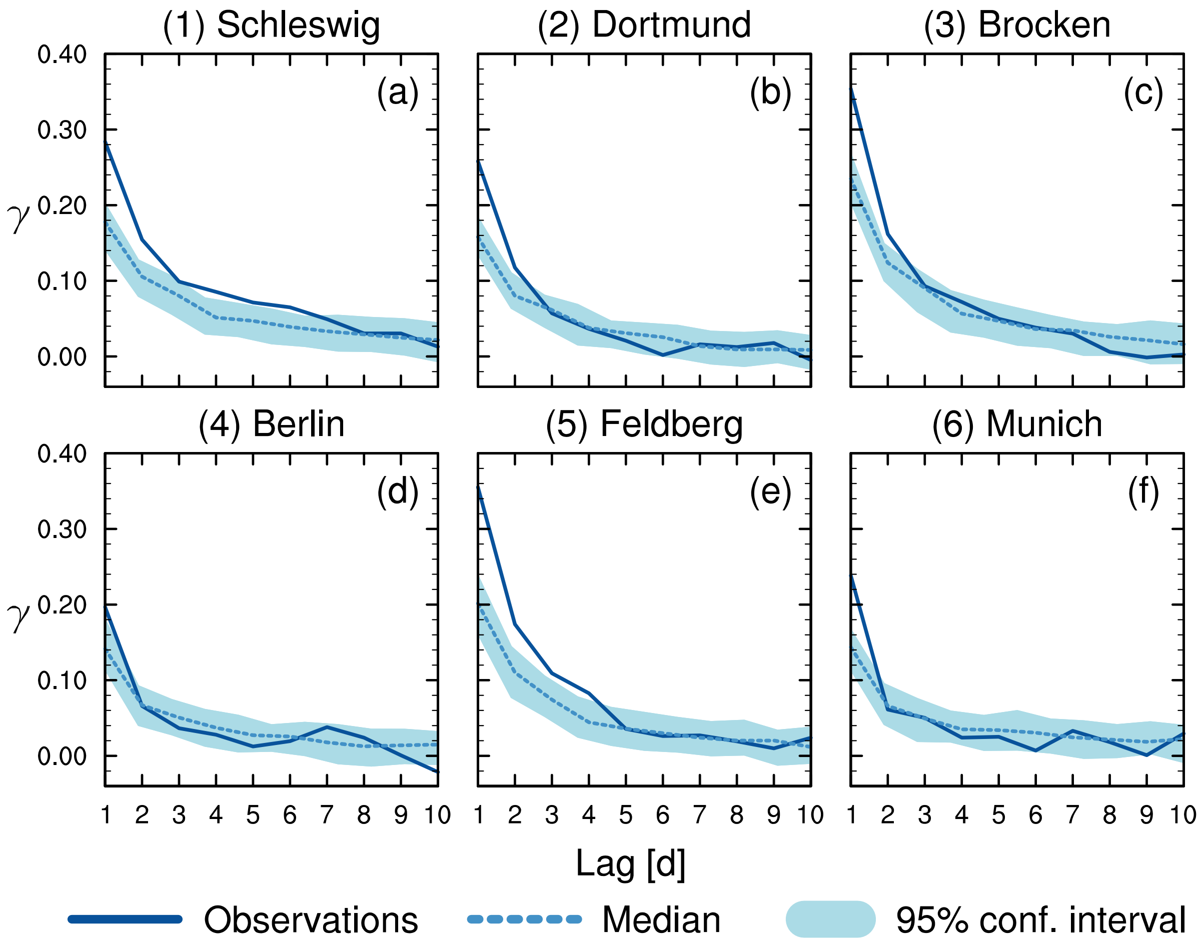


Figure 3.14: Autocorrelation γ for daily values up to ten lag days at six locations (Figure 3.3) during the evaluation period. Solid lines depict the autocorrelation for the observations. Dashed lines show the median autocorrelation for the simulations and the band quantifies the 95% confidence interval of simulations.

The coefficients γ obtained from the observations (solid blue line) decrease sharply with increasing lag time at all locations because the precipitation fields over Germany are governed by synoptic circulation patterns which exhibit a high temporal variability. The simulations also exhibit this feature, which is an emerging property of the cascade model structure (equation 3.1). For instance, if an eight day time step is dry, then also all corresponding daily time steps are modeled as dry.

Mountainous locations (3) and (5) exhibit the highest observed correlation coefficients γ at a one day lag time ($\gamma = 0.36$) (Figure 3.14c and 3.14e, respectively). In these locations, daily auto-correlation is high because precipitation amounts are strongly related to orography. The proposed WG is able to reproduce this feature to some extent because the fraction of dry time steps at each location is explicitly considered in the cdfs $\mathcal{F}$.

The auto-correlation of the generated time series are underestimating those of the observations at short lag times (i.e., less than three lag days), in particular at mountainous locations (Figures 3.14c and 3.14e). Explicitly introducing temporal correlations of the weights w in the proposed method may reduce this bias at the daily time scale. This, in turn, would at least double the computational cost of the proposed method because the lag-1 auto-covariance matrix among all locations has to be included which has the same dimension as a cross-covariance matrix $\mathcal{B}$.

In general, the temporal correlation is less than the spatial one for all locations (compare Figures 3.14 and 3.12). This fact justifies the assumption of the proposed method to explicitly consider the spatial cross-covariance (equation 3.3b) but neglect an explicit relationship to account for the auto-covariance of precipitation, which is frequently formulated in the literature (Wilks, 1998, 2009; Hundecha et al., 2009).

3.6 Conclusions

A novel method for seamless stochastic disaggregation of monthly to daily precipitation fields is proposed in this study. It is based on a multiplicative random cascade model that assumes quasi-stationary cdfs for partitioning precipitation instead of assuming stationary moments of precipitation. A modified sequential sampling algorithm using anchor points is introduced to ensure its applicability to spatial grids of any resolution and extent. Moreover, this method conserves the precipitation mass throughout multiple temporal scales, e.g., from 32 to 16 days and so forth until reaching daily values. The proposed method is tested on a high-resolution observational data set over Germany and complements spatial downscaling frameworks such as Wood et al. (2002) to provide a full spatio-temporal downscaling scheme for monthly data sets like Kirtman et al. (2014) and Harris et al. (2014).

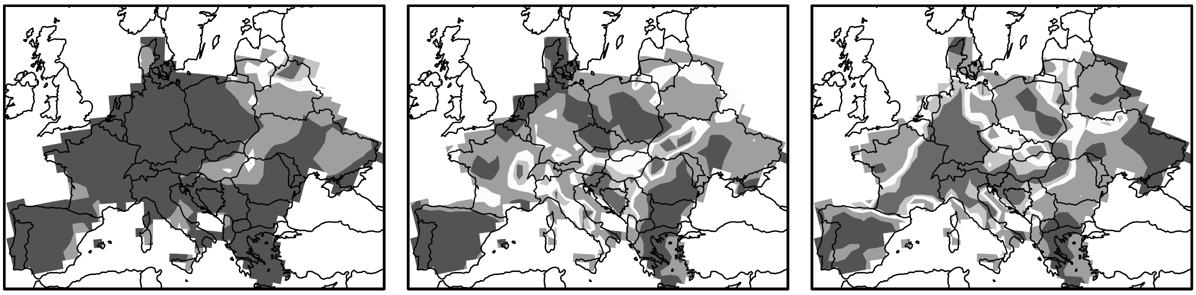
The analysis is carried out during a calibration and an evaluation period to test the assumptions of stationarity and evaluate model performance during a future period, which is frequently neglected (Charles et al., 1999). Results show that bias in precipitation distribution functions is negligible (<5%) for both the calibration and evaluation period for all considered time scales, which implies that distinctive location dependent precipitation characteristics are correctly estimated by the proposed WG. The particularly low mean absolute relative bias obtained during the evaluation period also highlights the capability of the proposed method during future periods.

Regarding extreme events, high precipitation amounts are correctly reproduced with bias less than 10% during the evaluation period for all temporal resolutions. The good agreement between the median simulated and observed spatial cross-covariance at different seasons and locations highlight the effectiveness of the spatial sampling algorithm. Moreover, the narrow confidence intervals of the simulations at daily and weekly temporal resolutions indicate the strength of the disaggregation mechanism. The temporal auto-covariance of daily precipitation is an emerging feature of the cascade model structure implemented in this WG, which is advantageous because no assumption in this respect is needed. These results highlight that the stationarity of the cdfs for partitioning precipitation holds quite well over Germany during the evaluation period, which is a crucial requirement for applications during future periods.

The proposed method employs very detailed information about the partitioning of precipitation at different locations, intensities, and temporal scales, which constitute an enhancement of available approaches such as Güntner et al. (2001). Further research is required to develop a parsimonious model that incorporates a suitable parametrization and regionalization of the cdfs $\mathcal{F}$. This regionalization should exploit the dependency between precipitation amounts and specific location characteristics such as elevation and geographic location. With respect to the cross-covariance matrices $\mathcal{B}$, it is worth investigating whether there are stochastic processes that would be able to approximate the empirical matrices sufficiently well.

Chapter 4

On the Capability of the North American Multi-Model Ensemble for Seasonal Soil Moisture Drought Prediction over Europe



Aknowledgments:
This study was carried out within the Helmholtz-Association climate initiative REKLIM (www.reklim.de) and was supported by the Helmholtz Interdisciplinary School for Environmental Research (HIGRADE). We acknowledge the E-OBS dataset from the EU-FP6 project ENSEMBLES (http://ensembles-eu.metoffice.com) and the data providers in the ECA&D project (http://www.ecad.eu). We also would like to thank the International Research Institute for Climate and Society (IRI; http://iridl.ldeo.columbia .NMME/) for making the NMME dataset available.

submitted as:
Stephan Thober, Rohini Kumar, Justin Sheffield, Juliane Mai, David Schäfer, and Luis Samaniego, "On the Capability of the North American Multi-Model Ensemble for Seasonal Soil Moisture Drought Prediction over Europe", Journal of Hydrometeorology, accepted with minor revision, 2015

4.1 Abstract

Agricultural droughts diminish crop yields and can lead to severe socio-economic damages and humanitarian crisis (e.g., famine). Hydrologic predictions of soil moisture (SM) that forecast droughts several months in advance are needed to mitigate the impact of these extreme events. In this study, the capability of a seasonal hydrologic prediction system for SM drought over Europe is investigated. The prediction system is based on meteorological forecasts of the North American Multi-Model Ensemble (NMME) that are used to drive the mesoscale Hydrologic Model (mHM). The skill of the NMME based forecasts is compared against those based on the Ensemble Streamflow Prediction (ESP) approach for the hindcast period of 1983-2009. The NMME based forecasts exhibit an Equitable Threat Score that is on average 69% higher than the ESP based ones at a six month lead time. Among the NMME based forecasts, the full ensemble outperforms the single best performing model CFSv2, as well as all subensembles. Subensembles, however, could be useful for operational forecasting at the expense of minor performance losses (less than 1%), but with substantially reduced computational costs (up to 60%). Regardless of the employed forecasting approach, there is considerable variability in the forecasting skill ranging up to 40% in space and time. High skill is observed when forecasts are mainly determined by initial hydrologic conditions. In general, the NMME based seasonal forecasting system is well suited for a seamless drought prediction system as it outperforms ESP based forecasts consistently over the entire study domain at all lead times.

4.2 Introduction

Droughts appear worldwide and belong to the most devastating natural catastrophes. Droughts are defined as dry anomalies and occur in all compartments of the hydrological cycle (Sheffield and Wood, 2011) such as the atmosphere (meteorological drought), streamflow and groundwater (hydrological drought), and root zone soil moisture (agricultural drought). We focus here on agricultural droughts because they are able to reduce crop yields leading to substantial socio-economic damages. For example, the 2003 European drought has caused losses of the order of 13 bn EUR (COPA-COGECA), whereas in the U.S. it is estimated that droughts lead to damages of 10 bn USD on average per event (Smith and Matthews, 2015). In developing countries, droughts even threaten the livelihood of societies. The 2010-2011 drought in the Horn of Africa, for example, led to a severe humanitarian crisis affecting around 12 million people (Relief, 2011; Dutra et al., 2013). Drought early warnings can help to mitigate the impact of these disasters several months in advance, but only if they are based on skillful seasonal forecasting systems.

State-of-the-art seasonal forecasting systems employ either dynamical or statistical frameworks to generate a drought forecast. Statistical frameworks, for example, use conditional distribution functions of observed historical datasets for drought prediction (Shahrbanou Madadgar and Hamid Moradkhani, 2013). Statistical frameworks, however, tend to be outperformed by dynamical ones, which represent the physics of the Earth system. Dynamical prediction systems typically constitute of coupled Global Circulation Models (CGCMs), which provide climate forecasts (CFs) of meteorological variables (e.g., precipitation and air temperature). These forecasts are then used to force a hydrological model that can reliably simulate the land surface components of the hydrological cycle such as root zone soil moisture (SM). Previous studies have assessed the forecast skill of experimental prediction systems for specific drought events (Luo and Wood, 2007; Dutra et al., 2013) as well as for multi-decadal hindcast periods (Shukla and Lettenmaier, 2011; Yuan et al., 2011; Wang et al., 2011; Mo et al., 2012; Yuan et al., 2013b,a; Mo and Lettenmaier, 2014; Shukla et al., 2014; Yuan et al., 2015). In these studies, the Ensemble Streamflow Prediction (ESP) approach is frequently used as a benchmark for representing climatological skill (Day, 1985). ESP is a statistical method that resamples meteorological forcings from a historic dataset to represent the forcing uncertainty under unknown future conditions. It has been used to discriminate between the impact of initial hydrologic conditions (IHCs) and that of CFs on hydrologic predictions (Wood and Lettenmaier, 2008; Shukla and Lettenmaier, 2011; Shukla et al., 2013).

Previous studies indicate that SM predictability depends strongly on the region considered. For example, ESP based SM forecasts in the Western United States are as skillful as CF based ones and the predictions only add value at one month lead time (Shukla and Lettenmaier, 2011; Mo et al., 2012). In contrast, the National Center for Environmental Prediction Climate Forecasting System (CFS) version one and two provide more skillful SM drought forecasts than ESP in the Central and Eastern United States up to six months lead time (Yuan et al., 2013b). This might be related to stronger correspondence of drought to the El Niño-Southern Oscillation (ENSO) in these regions and thus a higher atmospheric predictability (Mo, 2011; Mo and Lyon, 2015). A similar finding has been observed by Dutra et al. (2013) for a hindcast of the 2010-2011 Horn of Africa drought using the European Centre for Medium-Range Weather Forecasts (ECMWF) seasonal forecasting systems S3 and S4. They reported high predictability for periods associated with a La Niña event and less predictability otherwise. Although such ENSO teleconnections are weaker in Europe, Yuan et al. (2015) observed that CGCM based drought forecasts exhibit higher skill than ESP based ones up to five months lead times over the Danube river basin. In that study, the authors employed the recent North American Multi-Model Ensemble (NMME) which comprises 101 realizations of a multi-institutional, multi-model ensemble of climate forecast models up to lead times of 9-10 months (Kirtman et al., 2014). The spatio-temporal distribution of SM drought forecasting skill using NMME over Europe has, however, not yet been fully evaluated. A high forecasting skill irrespective of the location and lead time is a fundamental requirement for a seamless prediction system.

Few studies focused on drought predictability during particular drought phases such as the development, onset, and recovery. In one of these, Mo (2011) reported that drought recovery is more difficult to predict as it evolves on a shorter time scale than the development. Yuan and Wood (2013) reported that NMME models add skill to forecasts of meteorological drought onsets in tropical regions, but not in extra-tropical ones. In contrast to precipitation, SM drought predictability depends strongly on the IHCs (Wood and Lettenmaier, 2008), which are substantially drier during the recovery than during the development phase. This characteristic has not been exploited when investigating the impact of IHCs on SM forecasts.

Multi-model forecasting ensembles such as CFSv2, ECMWF S4, and NMME have ever-increasing ensemble sizes to provide a better estimate of model uncertainty. This implies that they also offer more than one meteorological forcing time series for assessment studies. Nonetheless, most assessment studies focus only on the grand ensemble mean (Dutra et al., 2013; Mo et al., 2012; Yuan et al., 2013b, 2015, among others). Few studies related the performance of the grand ensemble to that of individual models (Yuan and Wood, 2013; Mo and Lettenmaier, 2014). Thober and Samaniego (2014) recently showed that investigating subensembles, which do not take all realizations into account, has the potential to increase ensemble performance for reproducing extreme precipitation and temperature indices. Considering the fact that SM predictability is highly dependent on the quality of precipitation forecasts, subensembles could help either to increase the forecasting skill, or to reduce computational load for operational forecasts without loosing predictability.

Given the current knowledge regarding NMME based SM drought forecasts over Europe, four research questions constitute the main goal of this study. 1) Are NMME based drought forecasts more skillful than ESP based ones over larger parts of the European domain? 2) How is the drought forecasting skill distributed in space and time? 3) How skillful are subensembles in forecasting European droughts in comparison to single NMME models and the full ensemble? 4) How do IHCs impact drought forecasting skill during drought development and recovery?

To address these research questions, the mesoscale Hydrologic Model (mHM, Samaniego et al., 2010; Kumar et al., 2013a) is used to simulate SM for monthly NMME based precipitation and air temperature forecasts for the hindcast period of 1983-2009. These NMME based forecasts are contrasted against those based on the ESP approach, which serve as a benchmark in this study. The mHM derived SM forecasts are then transformed to a quantile based soil moisture index (SMI). The SMI lies in the interval [0,1] and a threshold of 0.2 is used to classify droughts. This cutoff implies that the lower 20% of SM states occurring in a given time period (e.g., a month) are considered as drought. Reference SMI fields are created using the observation based E-OBS dataset (Haylock et al., 2008) to assess the skill of the different forecasting approaches employing the Pearson correlation coefficient and the Equitable Threat Score (ETS).

4.3 Methods and Datasets

4.3.1 Climate Forecasts (CFs)

The forecasting dataset used in this study incorporates realizations of eight global climate models from the NMME with ensemble members varying between 6 and 24 per model (Table 4.1, see also Kirtman et al. (2014)). Monthly CFs of precipitation and air temperature are provided globally at a $1^\circ \times 1^\circ$ spatial resolution for lead times up to eight months. In total 101 realizations are used in this study available from the International Research Institute for Climate and Society. The performance of these models for SM drought forecasts is analyzed for the overlapping hindcast period of 1983-2009. The analysis is conducted over the European domain covering an area between 10°W-45°E and 35°N-55°N.

Acronym	Model	Institute	Ensemble members
CCSM3	Community Climate System Model, Version 3	University of Miami, Rosenstiel School of Marine and Atmospheric Science	6
CM2p1	Climate model version 2.1	Geophysical Fluid Dynamics Laboratory	10
ECHAMA	ECHAM version 4.5 anomaly coupled	International Research Institute for Climate and Society	12
ECHAMD	ECHAM version 4.5 direct coupled	International Research Institute for Climate and Society	12
GEOS5	Goddard Earth Observing System Model version 5	National Aeronautics and Space Administration	12
CFSv1	Climate Forecasting System version 1	National Center for Environmental Prediction	15
CFSv2	Climate Forecasting System version 2	National Center for Environmental Prediction	24
CanCM3	Canadian Coupled Global Climate Model version 3	Canadian Meteorological Center	10

Table 4.1: Climate Forecasting models used in this study, Institute they are developed at, and ensemble members available (see Kirtman et al. (2014) for details).

4.3.2 Construction of Soil Moisture Forecasts

The well-constrained mesoscale Hydrologic Model (mHM, Samaniego et al., 2010; Kumar et al., 2013a) is used here to generate gridded estimates of SM fields over the study domain. mHM is a spatially explicit distributed hydrologic model in which hydrological processes are conceptualized similar to these of other existing large-scale models like the VIC (Liang et al., 1996) and the WaterGAP model (Döll et al., 2003). It is driven by daily gridded fields of precipitation, air temperature, and potential evapotranspiration to simulate different components of the terrestrial hydrological system such as canopy interception, snow accumulation and melt, soil moisture and infiltration, runoff generation and evapotranspiration, deep percolation and base flow, and flood routing between grid cells. The model is open source (www.ufz.de/mhm) and readers interested in more details may refer to Samaniego et al. (2010). To date, mHM has been successfully applied to several river basins in Germany, North America, and Europe (Samaniego et al., 2010, 2013; Kumar et al., 2013a,b; Samaniego et al., 2014). In this study, a similar model setup with respect to terrain, soil, and land cover characteristics is used as in Rakovec et al. (2015), who demonstrated the ability of mHM to adequately represent the spatio-temporal dynamics of runoff, evapotranspiration, soil moisture, and total water storage anomaly over a wide range of European river basins.

The reference monthly SM field is obtained by forcing mHM with the observation based gridded E-OBS dataset (v8.0, Haylock et al., 2008) during the period 1950-2010. The E-OBS dataset is aggregated to 1° grid resolution to be compatible with the resolution of the NMME dataset. This reference SM field is then used to represent IHCs at the beginning of each month during the hindcast period (1983-2009).

Furthermore, the E-OBS dataset is used to set up the NMME and ESP based forecasts. The ESP forecast ensemble is created by resampling the meteorological dataset (i.e., E-OBS) of the hindcast period for a given target month excluding the year of that month, which is similar to the approach of previous studies (Twedt et al., 1977; Day, 1985; Wood and Lettenmaier, 2008; Shukla et al., 2013, among others). In total, the ESP forecasting ensemble consists of 26 members. The spatio-temporal variability of the E-OBS dataset is employed to disaggregate NMME based monthly precipitation forecasts to their corresponding daily values using a multiplicative cascade approach (Thober et al., 2014). This approach preserves the observed spatial patterns at the daily time scale as well as the monthly amount of the forecasted precipitation. Each monthly NMME forecast is stochastically disaggregated to an ensemble of 25 daily realizations, thus increasing the overall ensemble size to 2525 (= 101 × 25). The daily weights for disaggregating the monthly temperature forecasts are derived from the E-OBS dataset for a given target month. This procedure is similar to the rescaling technique used by Yuan et al. (2015). The rescaled temperature estimates are then also used to adjust potential evapotranspiration, which is calculated using the Hargreaves-Samani approach (Hargreaves and Samani, 1985). The daily mHM derived SM fields for both forecasting systems are then averaged to their monthly estimates. A representative SM field for a given NMME model realization is created by averaging the corresponding estimates derived from the 25 disaggregated meteorological forecasts because there is no significant variability among the latter fields as they are all forced with the same monthly precipitation and air temperature.

4.3.3 Calculation of Soil Moisture Index

The monthly SM fields are converted into their respective quantiles using a non-parametric kernel density estimation method for the drought analysis. The kernel density $\hat{f}(x)$ is estimated by

$$\hat{f}(x) = \frac{1}{nh} \sum_{k=1}^{n} K\left(\frac{x - x_k}{h}\right) \tag{4.1}$$

for a given sample of n SM fractions $x_1, \ldots, x_n$, bandwidth h, and kernel function K. A Gaussian kernel is used in this study and the bandwidth h is estimated by an optimization against a cross-validation error estimate (see Samaniego et al. (2013) for details). The respective quantiles, hereafter denoted as soil moisture index (SMI), and the corresponding distribution functions are estimated for each grid cell and calendar month independently. This procedure removes the seasonality of simulated SM and allows the comparability of SMI across locations. A SMI threshold value of 0.2 is used here to identify drought events following previous studies (Andreadis et al., 2005; Vidal et al., 2010; Sheffield et al., 2012; Samaniego et al., 2013, among others).

The monthly SM estimates are converted to their respective standardized anomalies prior to the conversion of SM to SMI to ensure their comparability across different realizations, climate models, and forecasting methods (Koster et al., 2009). The standardized anomalies are obtained by removing the seasonal mean and standard deviation. In this approach, the distribution function $\hat{f}$ is estimated only once using the reference SM anomalies. The forecasted SM anomalies are converted to SMI using this unique distribution function. This procedure provides a fair comparison between NNME and ESP based forecasts. In this study, no bias correction is applied to the NMME forecasts because the SMI calculation and the standardization of SM forecasts accounts for biases, particularly in the mean and standard deviation. The standardization of SM has also been exploited in previous studies to ensure comparability among different SM products (Dirmeyer et al., 2004; Koster et al., 2009; Wang et al., 2011). It is worth mentioning that bias correction is crucial for the correct quantification of hydrological fluxes in other applications such as streamflow predictions (e.g., Luo et al., 2007; Mo and Lettenmaier, 2014).

Three SMI forecasting ensembles are created in this study: two based on NMME forecasts and one based on ESP. The two NMME based approaches differ with respect to the employed averaging scheme. In the first approach, SMI forecasts are created for all 101 model realizations independently and these are then averaged to obtain a grand NMME ensemble mean for SMI. This approach is denoted as $\overline{\mathrm{SMI}}$. In the second approach, the SM fields are first averaged over all model realizations to create a grand NMME ensemble mean for SM. The latter is then transformed to its respective SMI. This approach is denoted as $\mathrm{SMI}(\overline{\mathrm{SM}})$. These two approaches will provide different results, because the SMI calculation is a highly non-linear transformation. Investigating these two averaging schemes will help to determine the best possible NMME drought forecasting skill.

4.3.4 Subensemble Selection

The NMME based forecasts are further evaluated with respect to the performance of subensembles, as these might give a better performance as the full ensemble but with a reduced computational demand. The backward search algorithm suggested by Thober and Samaniego (2014) is used to identify the best performing subensemble. This algorithm is computationally efficient because it does not require the evaluation of all possible subensemble combinations. The algorithm is summarized here:

1. Select all NMME models as the first subensemble.
2. Sequentially remove a remaining model from the subensemble and evaluate the corresponding performance (e.g., Pearson correlation coefficient R).
3. Repeat step 2 for all remaining models contained in the subensemble.
4. Replace the subensemble with the combination exhibiting the highest performance found in steps 2 and 3.
5. Repeat steps 2 to 4 until the subensemble contains only a single model.
6. Select the combination with the highest performance as the best performing subensemble.

4.4 Results and Discussion

4.4.1 Representation of Spatio-Temporal SMI Dynamics and the Effect of Model Averaging

The overall skill of the NNME and ESP based forecasts to mimic the spatio-temporal dynamics of the reference SMI is analyzed for different lead times using the Pearson correlation coefficient R (Figure 4.1). Two different averaging schemes have been employed to create the NMME based forecasts

(Section 4.3.3). All three methods have a comparably high skill at one month lead time (R ≈ 0.9), confirming the strong influence of IHCs on SM forecasts at a short lead time (Wood and Lettenmaier, 2008; Shukla et al., 2013). Expectedly, the forecasting skill decreases with increasing lead time, but the rate of this decrement is method dependent. For instance, the spatially averaged R value for ESP based forecasts drops from 0.90 at one month lead time to 0.32 at six months lead time (around 65% loss; Figure 4.1, panels g-i). For NMME based forecasts, which have been created by the $\overline{\text{SMI}}$ averaging approach, the skill decreases from 0.87 to 0.25 (around 71% loss; Figure 4.1, panels a-c). This is the strongest decrement among all considered methods, and also the lowest performance at any lead time. On the contrary, NMME based forecasts created by the SMI($\overline{\text{SM}}$) averaging approach, have the highest performance and the lowest decrement among all considered methods (around 42% loss; Figure 4.1, panels d-f).

Although the different forecasting methods yield distinctively different skill, the spatial patterns among the corresponding forecasts are very similar (Figure 4.1, panels a-i). This is observed for any lead time. Regions exhibiting consistently higher skill are located for all methods in Poland, Northern France, and Eastern Ukraine and relatively less skill in the Alps (i.e., Northern Italy, Switzerland, and Austria) and in the Pyrenees along the Spanish-French border. These patterns compare remarkably well with those of the persistence map of reference SMI (Figure 4.1, panels j-l). A high persistence (i.e., auto-correlation) of reference SMI indicates that SM states are exhibiting a long memory, which induces a high dependence of SMI forecasts on IHCs. In this study, perfect knowledge of IHCs is assumed (i.e., they are the same for all forecasts and the reference dataset), which leads to a high SMI forecasting skill (i.e., a high R) at locations exhibiting high SM persistence. On the contrary, SMI forecasts at locations having a short memory will be more dependent on CFs and the large uncertainty therein reduces the ability to represent reference SMI dynamics. These results illustrate the complex interactions between IHCs, CF, and SMI forecasting skill.

Additional to the initial land surface conditions, the averaging scheme employed to create the NMME based forecast has a decisive impact on the skill of representing reference SMI dynamics (Figure 4.1, panels a-f). Notably, the ensembles created by the SMI($\overline{\text{SM}}$) averaging scheme outperform ESP based forecasts, while the ensembles created with the $\overline{\text{SMI}}$ approach do not. This implies that the kind of averaging applied can have large impacts on the conclusions drawn in previous studies investigating the capabilities of ensemble drought prediction systems (Wang et al., 2011; Mo et al., 2012; Mo and Lettenmaier, 2014; Yuan et al., 2013b, 2015). The SMI values of individual models are often recasted to the one of the ensemble in these studies and the skill of drought prediction systems might be further increased by using averaging schemes that preserve the frequency of SMI values and therefore capture extremes.

The impact of different averaging schemes on SMI dynamics is exemplary illustrated using the 24 member CFSv2 ensemble (Figure 4.2). A strong annual cycle can be observed for both the forecasted and the reference SM fractions. The mean SM forecast tends to overestimate the reference one, but the latter is mostly within the uncertainty bound of the forecast (Figure 4.2a). The SMI, however, does not exhibit an annual cycle because the climatology of SM is treated separately for each calendar month in the SMI estimation (Section 4.3.3). The ensemble SMI forecasts tend to show a similar temporal dynamic as the reference one, but at the expense of an increased model spread compared to their respective SM forecasts (Figure 4.2b). Due to the increased model spread for SMI, there

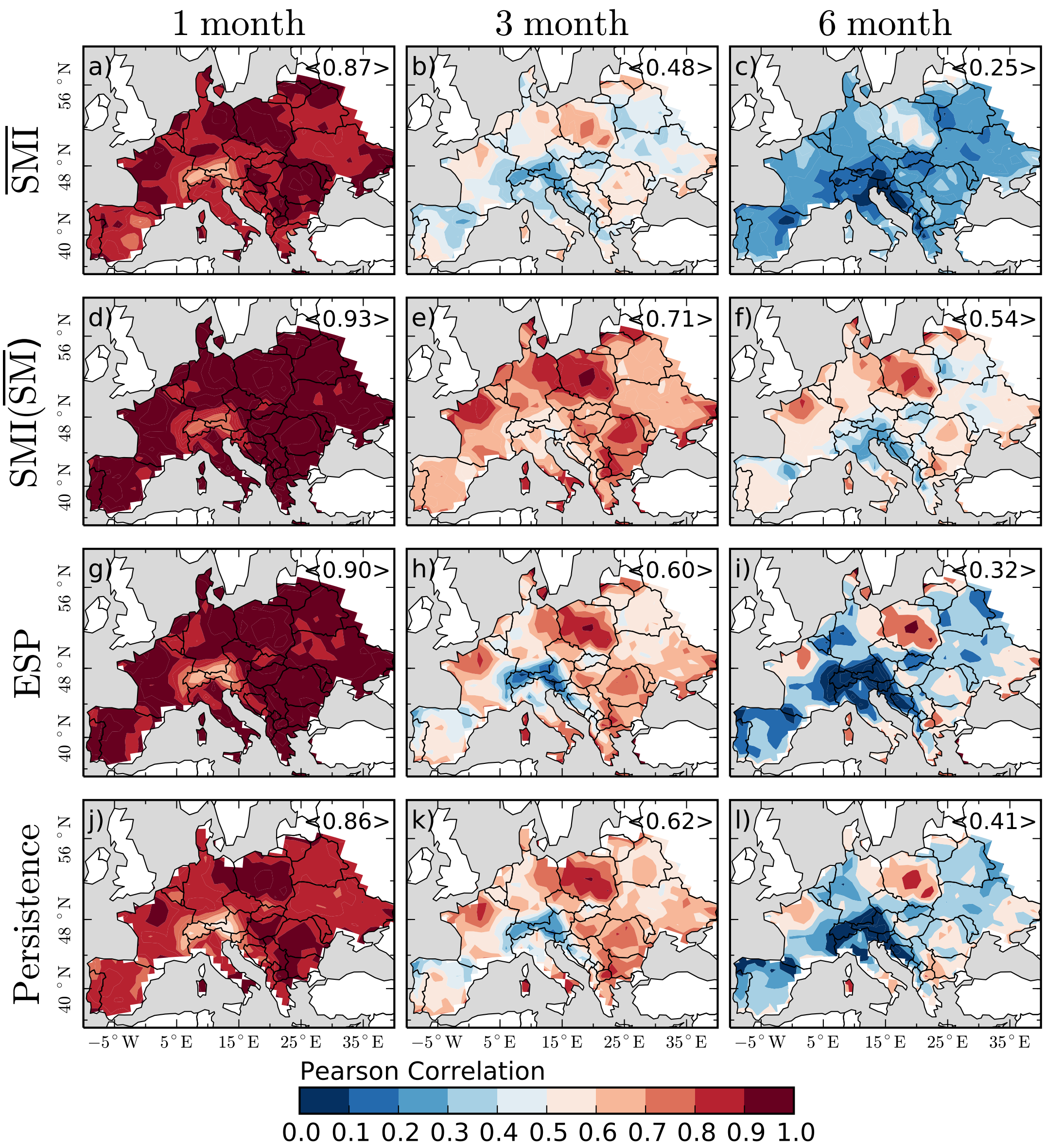


Figure 4.1: The skill to reproduce reference SMI is illustrated in terms of the Pearson correlation coefficient R between the forecasted and reference SMI for lead times of one, three, and six months. The skill of the NMME ensemble is depicted for two averaging schemes: $\overline{\text{SMI}}$ and $\text{SMI}(\overline{\text{SM}})$ in panels a-c and d-f, respectively. In the panels g-i, the skill of the ESP approach is shown. The persistence of reference SMI (estimated as Pearson auto-correlation) is displayed in the panels j-l. The spatial average of the corresponding R is depicted in the upper right corner of each panel.

is always a SMI forecast which is not under drought at a given forecast date. As a result, the $\overline{\text{SMI}}$ averaging approach does not detect drought events given a 0.2 drought threshold (i.e., no time step is identified as being in drought). The $\text{SMI}(\overline{\text{SM}})$ scheme captures both the wet and dry extremes better than the $\overline{\text{SMI}}$ scheme and also preserves the property that 20% of the SMI time steps are below

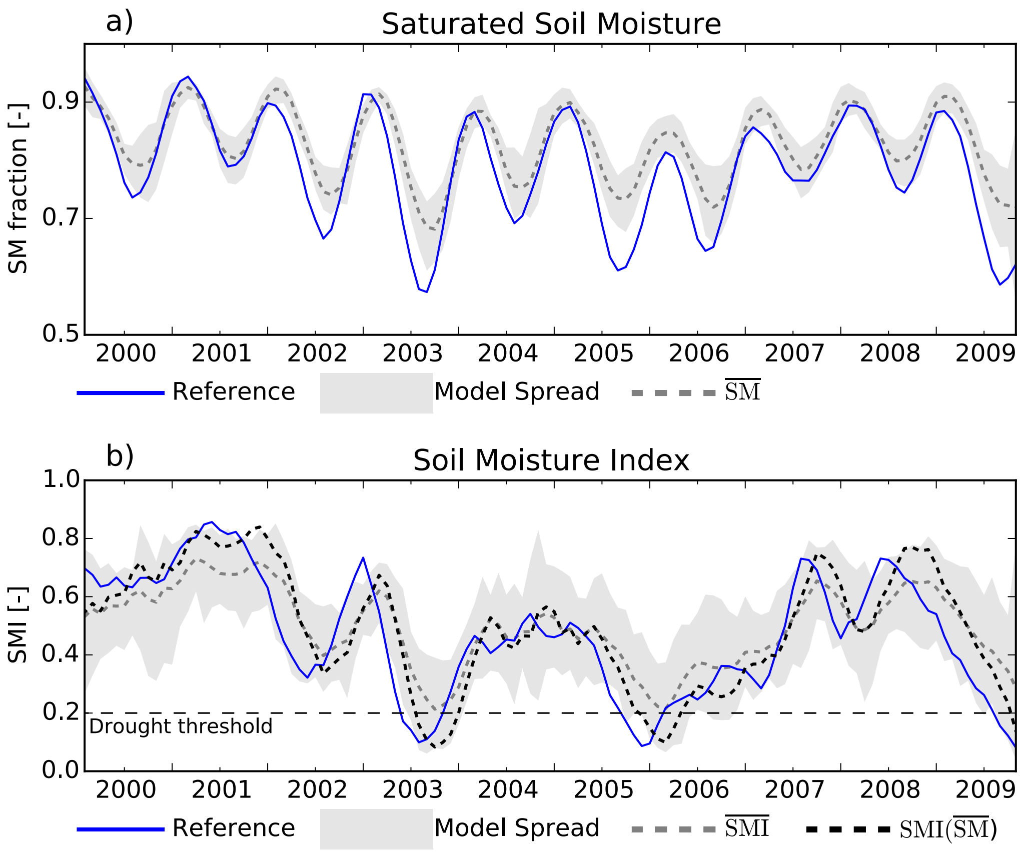


Figure 4.2: For a given grid cell (located in Central France at 47.19° N, 3.21° E), the exemplary time series of SM and SMI are depicted in panels a and b, respectively. In both panels, the blue line delineates the dynamics of the reference dataset and the gray band shows the uncertainty obtained from the 24 ensemble members of the CFSv2 forecasts at two months lead time. The gray dashed line in the top panel a denotes the average of the CFSv2 SM ensemble. The gray and black dashed lines in the bottom panel b denote the SMI ensemble derived by the $\overline{\text{SMI}}$ and SMI($\overline{\text{SM}}$) averaging scheme, respectively. The thin horizontal dashed line illustrating the drought threshold 0.2 is displayed for clarity.

0.2, which is crucial for drought analysis. The same effect was noticed for the other NMME models. Hence, the averaging scheme based on the SMI($\overline{\text{SM}}$) approach is used in the further analysis.

In general, the NMME based forecasts outperform the ESP based ones by 69% on average at a six month lead time (compare Figure 4.1, panels f-i). A similar outperformance has also been reported by Yuan et al. (2015) using bias corrected CFs. No bias correction is applied to the CFs in the present study because the SMI calculation using standardized SM anomalies already accounts for biases in the mean and standard deviation. This illustrates that bias correction of CFs might not be required to obtain a high forecasting skill for SM drought prediction. An analogous finding was reported by Yuan and Wood (2012) for streamflow, who demonstrated that driving a hydrologic model with raw CFs and subsequently bias correcting the simulated streamflow results in a skillful prediction of the latter.

4.4.2 Subensemble and Single Model performance for SMI and Drought Forecasts

Investigating the performance of subensembles is crucial to correctly determine the best possible performance of a given ensemble dataset. The backward selection algorithm proposed by Thober and Samaniego (2014) is used to identify subensembles of decreasing size based on Pearson correlation coefficient R and Equitable Threat Score (ETS), separately. The former criteria accounts for both wet and dry extremes, while the latter ETS is used to measure the skill of forecasts to capture drought events based on a 0.2 SMI threshold (see Appendix D for further details of the ETS). The performance criteria are averaged over space, lead time, and forecasting time step such that the selected subensemble exhibits a high skill regardless of location and time step considered, which is a basic requirement for a seamless prediction system.

The skill of any considered subensemble is higher than those of the single models for both criteria (Figure 4.3a). On the contrary, ESP has the lowest performance among all considered approaches for R and only marginally outperforms the worst performing model (CCSM3) for ETS. CFSv2 is the best performing model and the ordering of the single models is the same for R and ETS with the exception of the 2nd and 3rd best models which swap their places (CanCM3 and GEOS5). As a consequence, the models selected within the subensembles are quite similar for the two criteria (Figure 4.3b). Only the selected subensembles of size six are different by more than one model. For both criteria, the backward search algorithm correctly identifies CFSv2 as the single best performing model. It is worth noting that the algorithm would select a different model if the best performing model would have been deselected in a previous iteration. Such a result has been reported for the ENSEMBLES dataset (Thober and Samaniego, 2014).

The performance of the subensembles decreases monotonically with decreasing ensemble size for both criteria (Figure 4.3a). This justifies the approach pursued in previous studies to use the full ensemble as it exhibits the best possible performance (Yuan and Wood, 2013; Mo and Lettenmaier, 2014; Yuan et al., 2015, among others). However, the selected subensembles containing four models require 60% of the computational costs of the full ensemble to achieve a skill, which is only 0.3% and 0.5% less than that of the full ensemble for R and ETS, respectively. Operational forecasting could benefit from the reduced computational demand by using subensembles in favor of the full ensemble. The performance of the full NMME ensemble (NMME8) is, therefore, contrasted with that of a subensemble containing four models (NMME4) in the further analysis. Without loss of generality, NMME4 evaluated against ETS is chosen because it shows a similar performance as that evaluated against R (R value is only 1% less). The four models contained in NMME4 are CFSv2, CanCM3, ECHAMD, and CFSv1 (Figure 4.3b).

Although subensembles consistently outperform single models and ESP, the spread of both criteria is relatively narrow. This is due to the fact that the IHCs are the same for all forecasting methods, which reduces the variability among the different SM forecasts. In other words, the high variability in CFs is dampened while propagating through the hydrologic system exhibiting long memory. It is worth mentioning that substantially different subensemble performances have been observed for atmospheric variables like extreme precipitation indices (Thober and Samaniego, 2014).

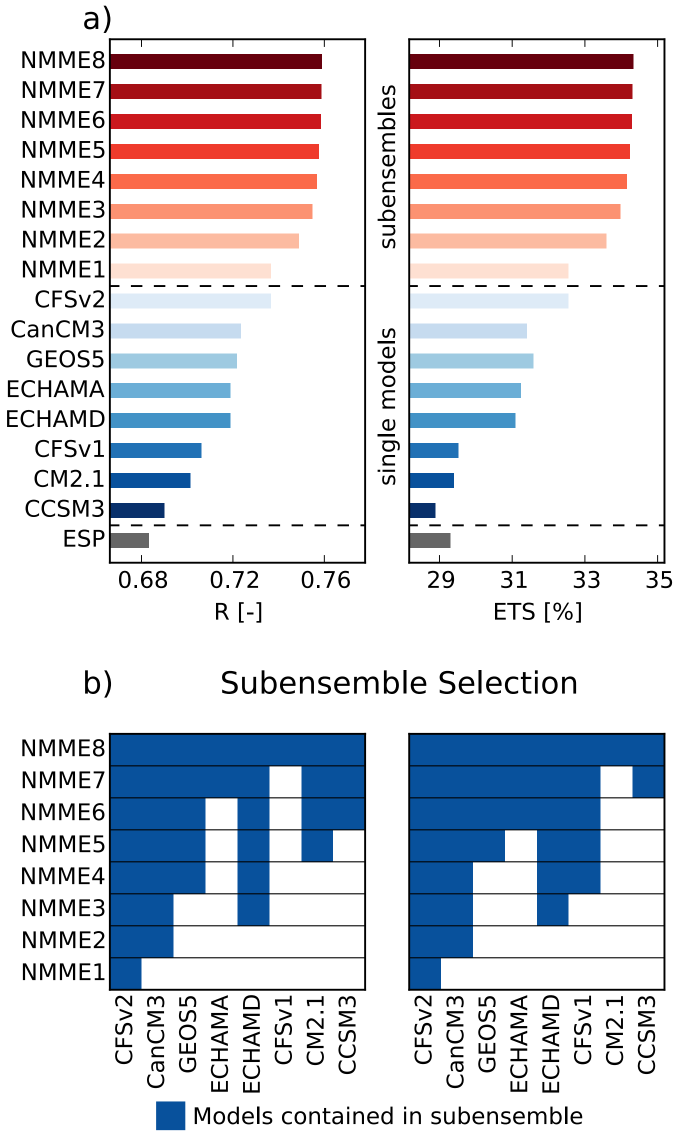


Figure 4.3: In the top panel a, the overall Pearson correlation and ETS estimates are shown for the NMME subensembles (red bars), single models (blue bars), and ESP (gray bar). These estimates are averaged over space and lead times to meet the requirements of a seamless prediction system. The SMI of NMME subensembles is obtained by the $\mathrm{SMI}(\overline{\mathrm{SM}})$ averaging scheme. In the bottom panel b, the single models contained within a selected subensemble for Pearson correlation and ETS are depicted by blue boxes.

4.4.3 Spatio-Temporal Distribution of Drought Forecasting Skill

It is desirable for a drought prediction system to be seamless with a high forecasting skill regardless of the location and the lead time. The forecasting skill of most prediction systems, however, varies in space and time (Shukla et al., 2013; Dutra et al., 2013; Yuan and Wood, 2013). The spatio-temporal distribution of ETS is analyzed here to understand these variations as well as the factors that influence drought forecasting skill.

Distinctive spatial patterns in ETS are observed for both NMME and ESP (Figure 4.4), which are similar to those of the Pearson correlation for the reference SMI dynamics (Figure 4.1, panels j-l). This illustrates that the impact of IHCs is also evident for extreme conditions. The differences in ETS between two locations across the study domain are as high as 40% (e.g., difference between Switzerland and Poland at one month lead time for NMME8; Figure 4.4a). These spatial differences are larger than the differences between the NMME8 and ESP forecasting approaches, which range up to 8% on average at six month lead time. It is worth noting that the spatial distribution between NMME8 and NMME4 is very similar (Figure 4.4). At 90% of the grid cells, the differences between these two ensemble based forecasts are smaller than 5% in terms of ETS irrespective of the lead time.

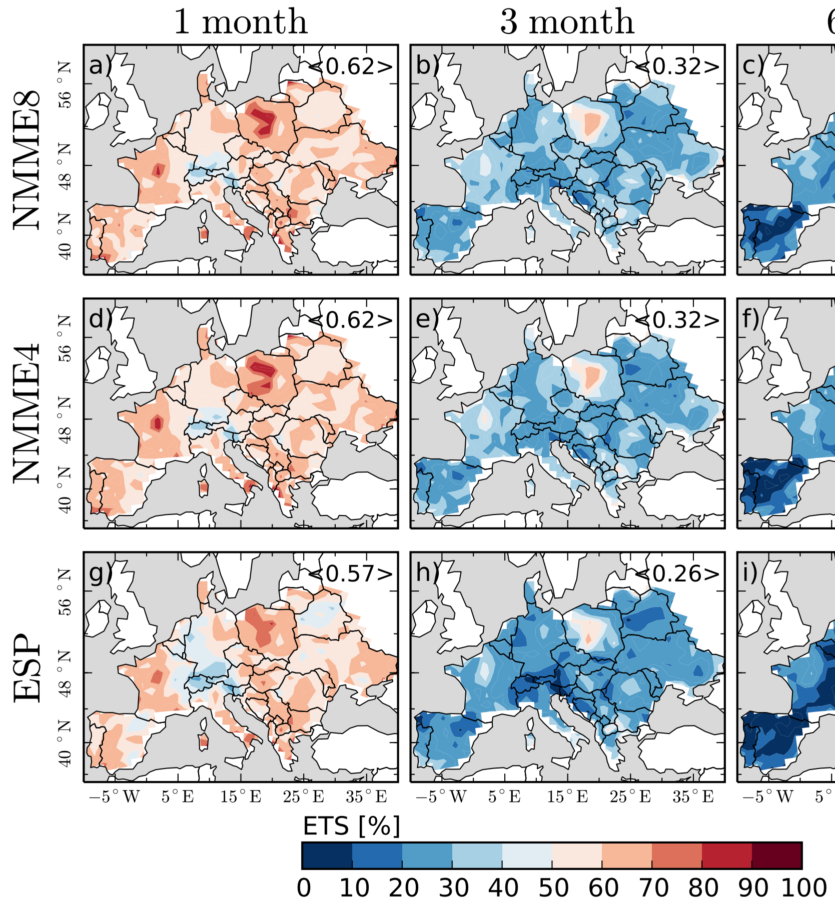


Figure 4.4: Spatial distribution of ETS at one, three, and six month lead time is displayed for the full NMME ensemble (NMME8) in panels a-c, NMME subensemble containing four models (NMME4) in panels d-f, and ESP in panels g-i. The NMME based forecasts are obtained by the $\mathrm{SMI}(\overline{\mathrm{SM}})$ averaging scheme. The corresponding spatial averages of ETS are denoted in the upper right corner of every panel.

This skill of both the NMME8 and ESP forecasting methods also depends on the forecast date (Figure 4.5, panels b-d). The differences between the smallest and highest ETS can be also as high as 40% for both forecasting methods, whereas the maximum difference between NMME8 and ESP forecasts at any given time step is at most 20%. Both forecasting methods, as expected, show lower ETS values at longer lead times, but the rate of decrement is less for NMME8 than for ESP. This leads to the relative outperformance of 69% on average at a six month lead time as discussed above (Section 4.4.1). These results illustrate the added value of an ensemble seasonal forecasting system at longer lead times (Mo and Lettenmaier, 2014). In general, NMME8 forecasts significantly outperform ESP ones at any location and lead time at a 5% significance level, which has also been reported by Yuan et al. (2015) using the VIC land surface model over the Danube basin in Europe. This result is obtained by applying a Student's t-test, which has been previously used in drought prediction studies (Wilks, 2011; Yuan et al., 2015). A similar result is obtained for the NMME4 subensemble, which requires only 60% of the computational demand as compared to the NMME8 (not shown).

The spread of single model performance is significantly narrower for the full NMME ensemble (19% on average) as compared to that of ESP (29% on average) at a 5% significance level (Figure 4.5, panels b-d). A similar result is obtained when the same number of samples (forcing members) is evaluated for NMME8 and ESP. The higher uncertainty for the ESP based forecasts can be mostly attributed to poorly performing forecasts. The spread of ETS for the NMME8 based forecasts is often located within the upper tail of that estimated for the ESP based ones. The skill of the full NMME ensemble is comparable to that of the best performing model at a given forecast date (i.e., the upper limit of single model spread shown in Figure 4.5, panels b-d), which has also been reported for an NMME based prediction system over the CONUS (Mo and Lettenmaier, 2014). It is worth noting that there exists not a single model that outperforms all others at all forecasting dates. For example, CFSv2 only outperforms all other models at 20% of all forecasting dates, although it is the overall best performing model (as discussed above; Figure 4.3). This again highlights the advantage of using ensemble based forecasts over ones based on a single model.

The temporal dynamics of ETS for the full NMME ensemble and ESP are quite similar (Figure 4.5, panels b-d), which again signifies the role of IHCs for drought predictions. Low ETS values are generally observed during periods of drought recovery with less extensive droughts (e.g., 1988, during autumn 1998, and at the end of 2004; Figure 4.5a). Both forecasting methods overestimate the drought extent during these periods, which results in a high false alarm rate and thus reduces ETS. On the contrary, high ETS values are observed during drought development phases (e.g., during 1990, 1994, and summer of 2005). These results illustrate that the drought forecasting skill varies depending on the states of drought events (e.g., drought development and recovery). These are defined in the following section.

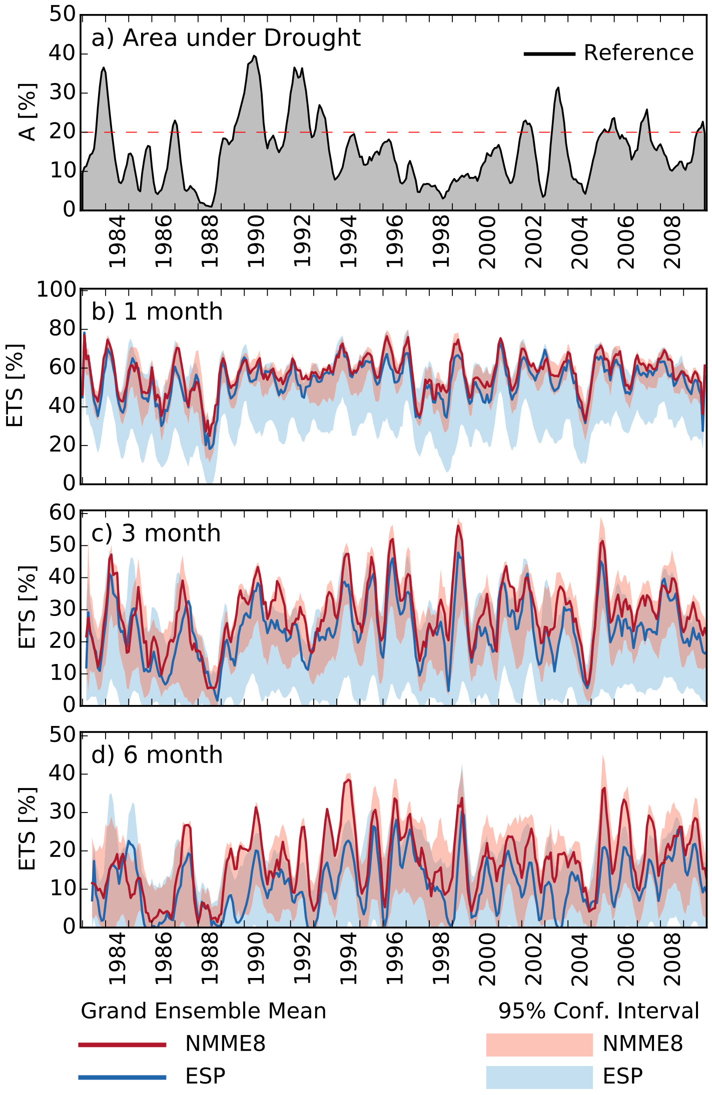


Figure 4.5: The top panel depicts the fraction of area under drought based on the reference SMI dataset. The thin horizontal dashed line is added for clarity displaying the threshold for droughts covering more than 20% of the European domain. Panels b-d illustrate the temporal variability of ETS for the ESP (blue lines) and the full NMME ensemble (red lines) based SM drought forecasts. The NMME based forecasts are obtained by the $\mathrm{SMI}(\overline{\mathrm{SM}})$ averaging scheme. Additionally, the 95% confidence interval for the single ESP and NMME ensemble members is depicted as light red and blue bands, respectively. Ticks mark the end of the respective year. The scale of y-axis are different for each panel for clarity.

4.4.4 Forecasting Skill during Drought Development and Recovery

To further investigate the forecasting skill during drought development and recovery phases, two drought characteristics are analyzed for major drought events that cover more than 20% of the European domain (e.g., the 1983, 1990, and 2003 drought; see also Figure 4.5a). A drought time step is defined as development (recovery) if it occurs before (after) the peak extent of the respective event. The two characteristics are the drought severity and the area under drought (see Appendix E for details). Both of these characteristics are normalized by their corresponding reference estimates (based on E-OBS) to make them comparable among different events. The perfect forecast would correspond to a value of one for both characteristics. The drought characteristics during both phases are calculated for all NMME and ESP ensemble members separately. Finally, a probability density function is estimated jointly for the two characteristics using a kernel estimation method (Equation 4.1) to assess their associated spread, following the procedure used by van Loon et al. (2014).

In general, the forecasted drought severity matches the median reference one quite well, with deviations less than 20% irrespective of the lead time, drought phase, and forecasting method (horizontal lines in Figure 4.6). On the contrary, substantial underestimations in drought area are observed with increasing lead time up to 55% for NMME8, 51% for NMME4, and 68% for ESP (vertical lines in Figure 4.6). Additionally, these are more pronounced during drought development phases than during recovery phases. In summary, the drought forecasts exhibit a higher mismatch in correctly detecting reference drought location. If a drought has been correctly forecasted at a given location, then it is likely that the severity of this event would be comparable to that of the reference one.

The spread of drought severity and area increases with lead time for all forecasting methods (see regions containing 90% of the density in Figure 4.6). Expectedly, the relatively larger uncertainty in CFs at longer lead times causes a higher spread in drought characteristics (Wood and Lettenmaier, 2008; Shukla and Lettenmaier, 2011). This spread is larger during the drought recovery than during the development phases at a long lead time, which is in agreement with Mo (2011) who reported that drought development is more predictable than drought recovery.

The spread is also remarkably similar for the NMME8 and NMME4 based forecasts. For example, there is a comparable overlap of spread estimated during the drought development and the recovery phases at a three month lead time. This overlap is considerably different from that observed for ESP based forecasts (Figure 4.6, compare panels b, e, and h). These results illustrate that the NMME4 subensemble also has a similar performance as the full NMME ensemble during different drought phases, but only requiring 60% of the computational resources.

In general, all forecasting methods underestimate the reference drought severity during the drought development phases at all lead times (Figure 4.6). This results from too wet forecasts leading to higher SM conditions as compared to the relatively drier reference ones. On the contrary, drought severity is overestimated during the drought recovery phases at three and six months lead times. The forecasts are drier than the reference one in this case. In other words, they are not able to add sufficient SM to recover from the drought. These results illustrate the fundamental influence of IHCs that persist throughout the drought forecasts leading to a consistent lag of these with respect to the

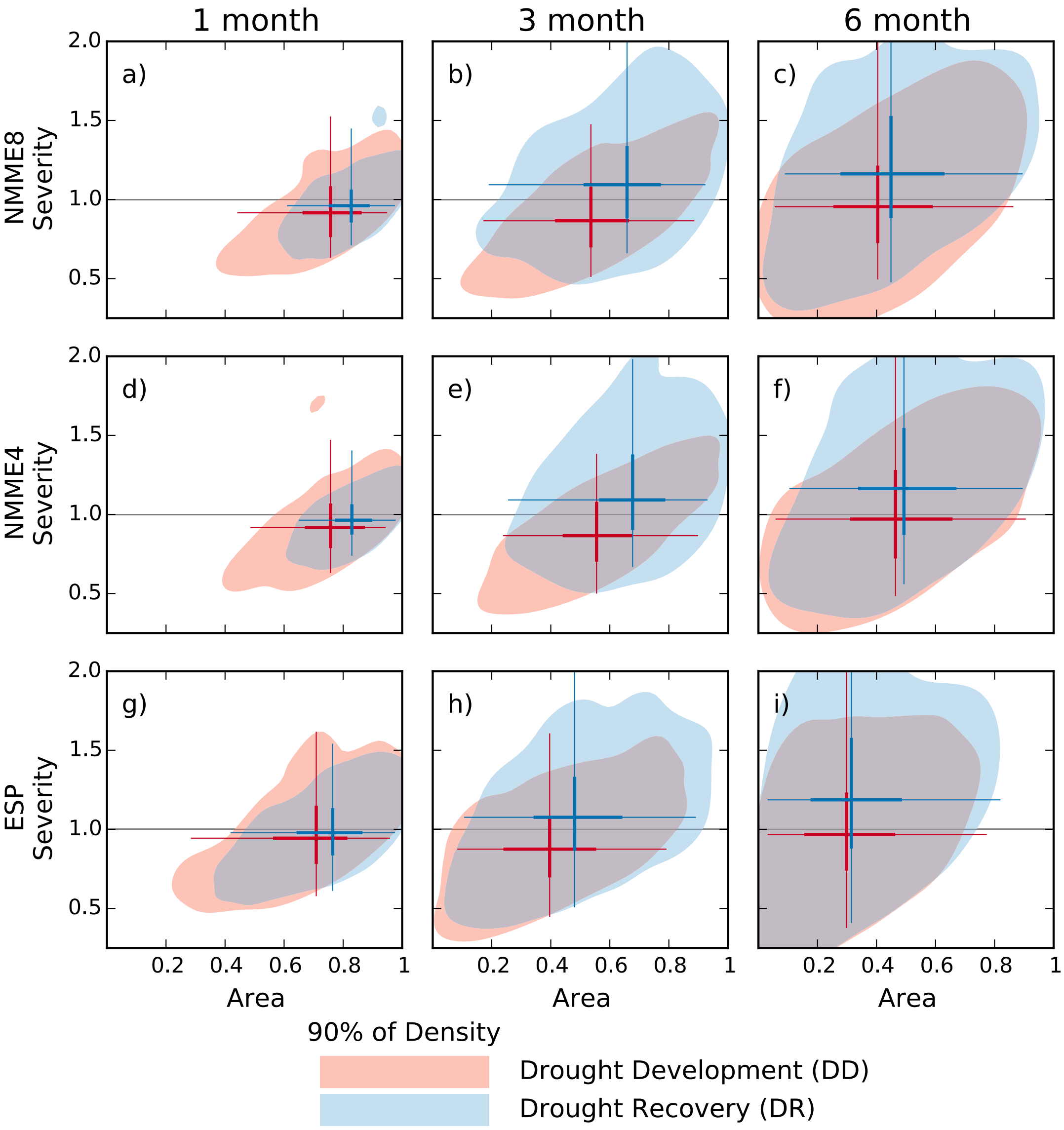


Figure 4.6: Probability density function for drought severity and drought area is illustrated for different lead times for forecasts obtained by single NMME and ESP ensemble members. The performance for all NMME models is shown in panels a-c and only for four NMME models in panels d-f. The performance for ESP based forecasts is displayed in panels g-i. The area containing 90% of the density for both characteristics is depicted in each panel as red and blue regions for drought development and recovery, respectively. Additionally, the spread for each characteristic is shown as box plots for the different drought phases (95% confidence interval as thin lines, the spread between the 25th and 75th quantiles as thick lines, and the median is located at the intersection).

reference SMI dynamics (see also Figure 4.2b). This is expected for ESP as it represents a climatological forecast and the skill is mainly derived from the correct representation of IHCs (Koster et al., 2004; Shukla et al., 2013). The skill of NMME based forecasts has a similar dependence on the IHCs as ESP despite that NMME models represent physical dynamics of the Earth system. They do, however, provide a substantially better forecast for drought area as compared to ESP (Figure 4.6).

4.5 Summary and Conclusions

In this study, the skill of a seasonal hydrologic prediction system for soil moisture (SM) drought forecasts is evaluated over Europe for a 27 year hindcast period (1983-2009). The prediction system is based on meteorological forecasts of the North American Multi-Model Ensemble (NMME) that are used to drive the mesoscale Hydrologic Model (mHM). The skill of NMME based forecasts is contrasted with that of the Ensemble Streamflow Prediction (ESP) approach. The obtained SM estimates from both forecasting approaches are transformed to a quantile based soil moisture index (SMI) to conduct a drought analysis using a 0.2 SMI threshold. Drought prediction skill is quantified in terms of the Equitable Threat Score (ETS) employing a reference SMI field. The latter has been created using the observation based E-OBS dataset.

NMME based forecasts significantly outperform ESP based ones particularly at a long lead time (i.e., up to 69% higher ETS at six month lead time). This is achieved only if the SMI has been calculated for the grand ensemble SM mean. In contrast, the grand ensemble SMI obtained by averaging single NMME model based SMIs does not outperform the ESP based one. Among the NMME based forecasts, the full ensemble outperforms the single models as well as all selected subensembles. There is a considerable variability in the skill of SMI forecasts over Europe (i.e., up to 40% in space and time), regardless of the forecasting approach. This variability is strongly related to the persistence of reference SM, illustrating the strong impact of initial hydrologic conditions (IHCs) on SM drought forecasts. The IHCs are respectively wetter during drought development phases than during drought recovery phases, which induces an underestimation of drought severity during the former and an overestimation during the latter phase.

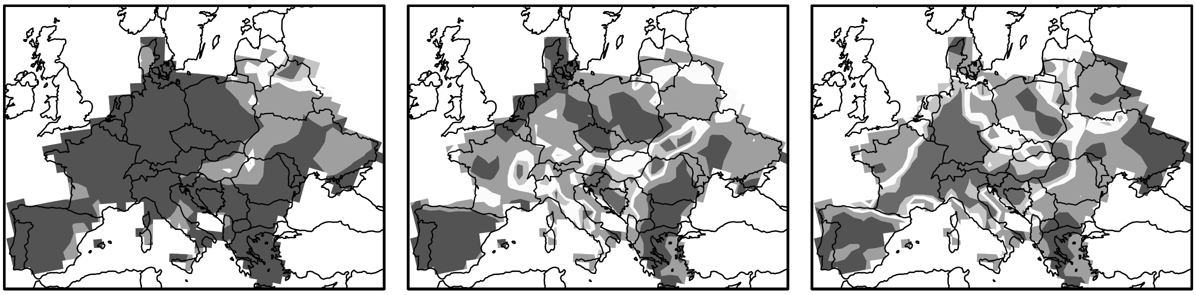
The main conclusion of this study is that NMME based forecasts are useful for seasonal SM drought prediction over Europe, which is in accordance with recent studies for the CONUS and GEWEX river basins using the VIC land surface scheme (Mo and Lettenmaier, 2014; Yuan et al., 2015). The NMME based forecasts are well suited for a seamless prediction system as their skill is consistently higher than that of ESP based ones over the entire study domain at all lead times.

Operational seasonal SM drought forecasting should consider using subensembles in favor of the full ensemble. The selected subensembles only show performance losses less than 1% on average in comparison to the full ensemble, but reduce the computational demand up to 60%. Moreover, bias correction of raw meteorological data has little impact on SM drought forecasting skill because the calculation of the quantile based SMI already accounts for systematic biases, particularly in the mean and standard deviation.

The results of this study illustrate the ubiquitous impact of IHCs on SM drought forecasting skill. The uncertainty associated with imperfect IHCs is, however, not considered here. Methods for further evaluating this aspect such as the reverse ESP approach have been investigated in previous studies using observational datasets (Wood and Lettenmaier, 2008; Shukla and Lettenmaier, 2011; Shukla et al., 2013). With the increase of computational resources, these should also be considered in the evaluation of ensemble SM drought prediction systems such as those based on the NMME.

Chapter 5

Discussion and Conclusions



This chapter first summarizes the results of the three preceding chapters and then presents a discussion on individual topics that relates these to each other and addresses limitations that should be investigated in follow-up studies. The conclusions of this discussion are presented in a final section.

5.1 Summary

The main focus of this work is the development and evaluation of a seasonal drought prediction system for Europe. The development has been separated into the investigation of subensemble selection methods to increase the performance of a given multi-model ensemble (Chapter 2) and the development of a novel approach to disaggregate monthly precipitation to daily values (Chapter 3). The methods developed in these two chapters have then been used in the design of a seasonal drought prediction system for Europe (Chapter 4). The following paragraphs outline the main results.

The evaluation of one of the latest collection of regional climate models over Europe provided by the ENSEMBLES project (van der Linden and Mitchell, 2009) showed that these models, expectedly, simulate extreme temperature indices better than extreme precipitation indices. With respect to precipitation, models tend to overestimate long-term annual precipitation amounts and underestimate the maximum number of consecutive dry days, which has also been observed in previous studies (Rauscher et al., 2010; Herrera et al., 2010). The spatial variability of extreme indices is also often underestimated by the considered regional climate models. In other words, the extreme indices obtained from regional climate models are too smooth in space leading to high discrepancies between simulated and observed extremes in complex terrain such as the Alps. It is worth mentioning that one model substantially outperformed all other models, which has also been reported for Portugal (Soares et al., 2012). As a consequence, this model even outperformed the full ensemble mean, which is frequently evaluated within climate change impact assessment studies (Giorgi and Mearns, 2002; Doblas Reyes et al., 2005; Weigel et al., 2010).

Different algorithms have been assessed in their ability to efficiently identify the highest performing subensembles. Among these, the newly proposed backward elimination algorithm was able to identify subensembles that exhibit a comparable performance as the best possible subensemble at a given ensemble size. In particular, it outperformed the weighted averaging approach (Giorgi and Mearns, 2002). The backward elimination method correctly identified the subensemble containing only six out of 13 regional climate models that gave the overall best possible performance, which outperformed all single models as well as the full ensemble mean.

The low resolution output of climate models needs to be downscaled to the resolution of hydrologic modeling to conduct a meaningful analysis. In this work, a newly proposed method for the stochastic disaggregation of monthly precipitation fields to daily ones is investigated. The method uses a multiplicative cascade approach with weights (i.e., multipliers) that are sampled consistently in space. A sequential sampling approach has been extended to be applicable to grids of any resolution and extent (Appendix A). The method has been tested on a high-resolution dataset over

Germany. It was shown that the employed statistical relationships are quasi-stationary in time making the method suitable for climate change investigations during future periods. The method consistently reproduced observed precipitation distribution functions at diverse locations and was also able to preserve extreme precipitation indices with biases less than 10%. It is worth mentioning that the multiplicative cascade model structure automatically created a realistic temporal auto-covariance of daily precipitation that compared well with the observed one without requiring any explicit assumption in this respect.

The seasonal drought prediction system utilized the multiplicative cascade approach investigated in Chapter 3 to disaggregate monthly precipitation forecasts by the North American Multi-Model Ensemble (NMME). The NMME-based forecasts exhibited a significantly higher forecasting skill in predicting agricultural droughts over Europe than the forecasts based on the Ensemble Streamflow Prediction approach, which has been used as a benchmark in this work (Chapter 4). The Equitable Threat Score, which summarizes the hit rate and false alarm rate into one measure of skill, has been employed to quantify the drought forecasting skill. The Equitable Threat Score was observed to be up to 69% higher at a six month lead time for the NMME-based forecasts. The results also showed that there is considerable spatial and temporal variability in drought forecasting skill. In space, drought forecasting skill is higher in regions where soil moisture itself exhibits a long memory. The perfect knowledge of initial hydrologic conditions then leads to a high forecasting skill in these regions. With respect to the temporal variability, low drought forecasting skill was observed during periods of little drought indicating that the forecasts overestimated drought extent during these periods. The backward elimination method investigated in Chapter 2 has also been used to analyze the performance of subensembles of these forecasts. The full ensemble mean exhibited the highest performance among all evaluated subensembles and also outperformed all single models. Subensembles that contained only 60% of the model realizations, however, exhibited an Equitable Threat Score that is only 1% less than that of the full ensemble mean. These might be highly beneficial for operational drought forecasting because of the substantially reduced computational demand at a moderate performance loss.

5.2 Discussion of Spatial Downscaling

The spatial resolution of the meteorological forcing dataset obtained from the North American Multi-Model Ensemble remains unchanged within the presented seasonal drought prediction system at a relatively coarse $1^\circ \times 1^\circ$ grid. The method developed in Chapter 3 was only used for the temporal disaggregation of monthly precipitation to daily values.

The sequential “Anchor sampling” method presented in Appendix A, which has been developed for the temporal disaggregation, could, however, also be useful for spatial downscaling because it is able to generate random fields with a prescribed spatial covariance structure at grids of any resolution and extent. For example, a method that is frequently used for the spatial downscaling of climate model output is quantile mapping. This procedure increases the spatial resolution of a meteorological variable by converting the quantiles of a low-resolution predictor to that of a high-resolution predictand. A quantile mapping algorithm essentially can be expressed by the following equation

$$x = F^{-1}(G(y)),$$

where x is the high-resolution predictand with the cumulative distribution function F and y the low-resolution predictor with distribution function G. The applicability of this method is currently under vivid scientific debate because it neglects the variability in the high-resolution predictand given a low-resolution predictor (von Storch, 1999; Maraun, 2013).

This variability can be visualized by a two dimensional histogram of the marginal distributions of the respective meteorological variable at a high and low spatial resolution. These histograms are expressing the copula, which describes the statistical dependency between the marginal distributions of two random variables (Nelsen, 2006). The copula can be generally written as

$$\mathbf{P}(x \leq X, y \leq Y) = C(F(x), G(y)),$$

where x, F, y, and G are defined as above, the left hand side of the equation is the joint probability $\mathbf{P}$ of x and y, and C is the copula between F and G. Figure 5.1 depicts the empirical copula between precipitation distribution functions at a 8 km and a 4 km resolution over entire Germany. This

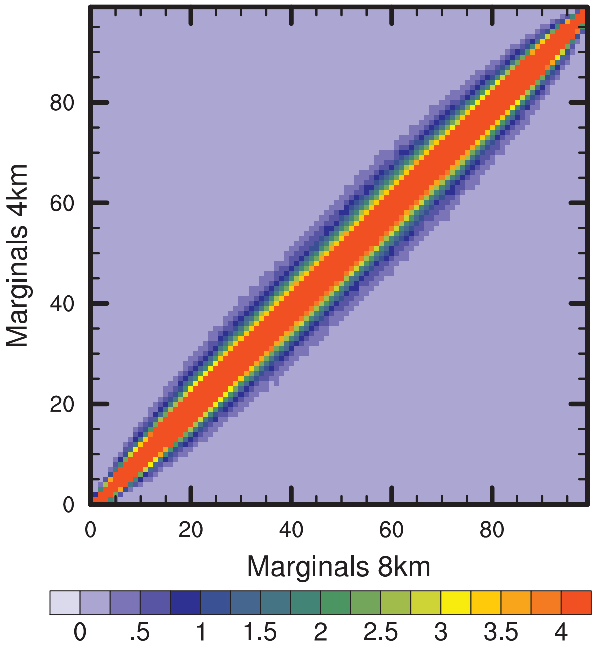


Figure 5.1: Two dimensional histogram of monthly precipitation distribution functions at a 8 km (x-axis) and 4 km (y-axis) spatial resolution. The bin width of the histogram is 1 percent in both distribution functions. The marginal distributions have been estimated as empirical distribution functions of the same dataset investigated in Chapter 3 (i.e., gridded precipitation dataset with a 4×4 km^2 spatial resolution over entire Germany available for the period from 1961 to 2010). This figure is identical to the empirical copula for the two quantities.

framework quantifies the conditional probability of observing a high-resolution value given a low-resolution one. It is highly advantageous in comparison to quantile mapping which neglects exactly the uncertainty that is expressed in this conditional probability and simply relates, for example, the median value of the predictor with the median of the predictand. Copula methods are, however, only rarely applied for the spatial downscaling of precipitation (van den Berg et al., 2011).

The disadvantage of a copula downscaling in contrast to a quantile mapping is that it requires to draw a random number to determine which percentile value is chosen from the conditional probability for the high-resolution predictand. If these random variates are determined independently at each grid cell, then the downscaled field will preserve the spatial distribution of the coarse scale predictor. But it is going to be unrealistically noisy within one low-resolution grid cell because no spatial covariance is imposed on the downscaled values. This is visualized in Figure 5.2 for the downscaling of the monthly precipitation for August 1992 from a 12 km spatial resolution to a 4 km one. While the high-resolution field preserves the overall spatial distribution of precipitation of the low-resolution one, it is very noisy, in particular in the region exhibiting high precipitation values in Northwestern Germany.

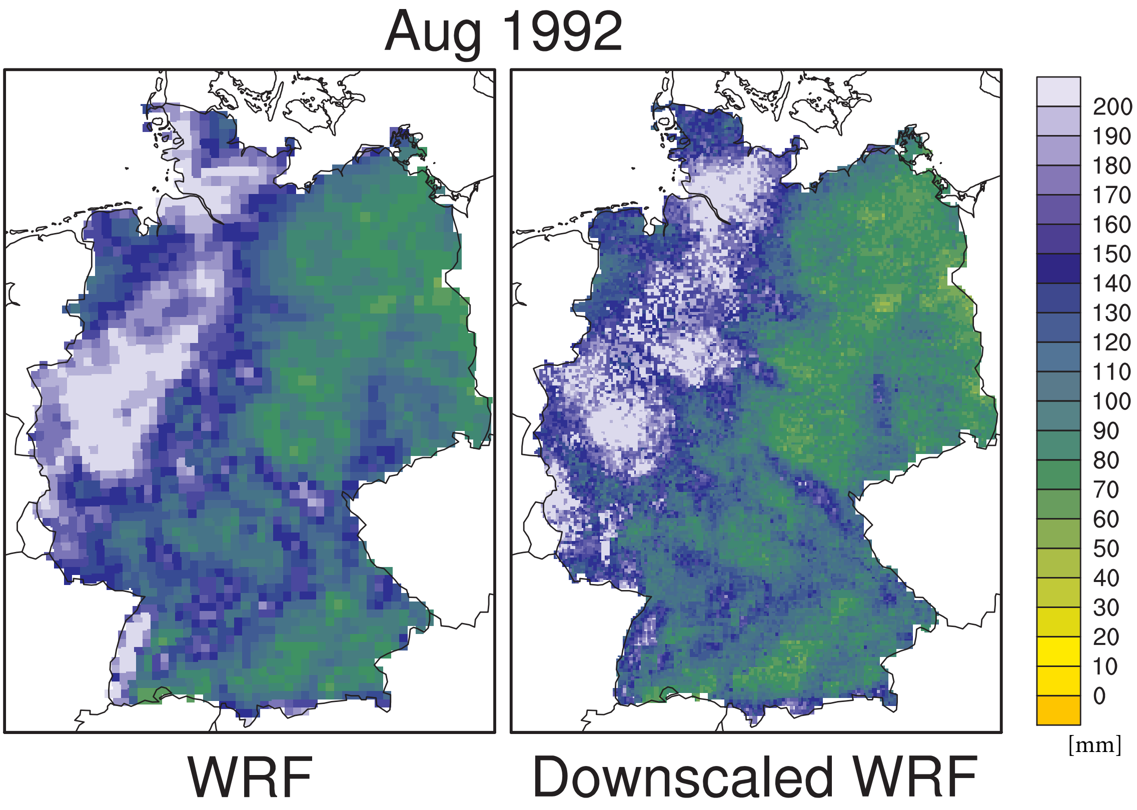


Figure 5.2: One example for a copula downscaling of WRF precipitation from a $12\times12\,\mathrm{km}^2$ to a $4\times4\,\mathrm{km}^2$ resolution. This downscaling is conducted for the monthly precipitation amount of August 1992. The WRF data has been kindly provided by Dr. Kirsten Warrach-Sagi (University of Hohenheim).

A spatial covariance structure thus has to be enforced on the downscaled percentiles. The newly proposed sequential sampling algorithm proposed in Appendix A ("Anchor Sampling") is fitting this purpose very well because it has been designed for large gridded datasets which are typically used within copula downscaling schemes. It can be used to sample random variates with a predefined covariance structure that is matching the observed one quite well. The applicability of the sequential sampling method is not limited to temporal disaggregation or copula downscaling schemes. This method could be beneficial for any statistical downscaling method that relies on the sampling of random variates with a predefined covariance to create local scale variability.

The statistical relationship of a meteorological variable at two different resolutions is also dependent on the temporal resolution. For example, the relationship depicted in Figure 5.1 is substantially different for daily values because of the occurrence of zero precipitation amounts (not shown). These lead to a spatial intermittency of precipitation that has to be introduced by the spatial downscaling procedure. This intermittency is typically less pronounced at lower temporal resolutions such as monthly values (also seen in Figure 3.6) making the spatial downscaling relationships less complex at these resolutions. The newly proposed stochastic disaggregation method has proven to be able to introduce a realistic spatial intermittency as the temporal resolution is increased from monthly to daily values. The method thus allows to first perform a spatial downscaling at a relatively coarse temporal resolution such as monthly values and to subsequently increase the temporal resolution to daily values (Figure 1.4).

5.3 Discussion of Subensemble Selection

Different ensemble selection algorithms have been investigated in Chapter 2. The different selection methods have been tested for regional climate models over Germany with a focus on extreme precipitation and temperature indices, which are important for hydrologic modeling. The backward elimination method has been identified as the most efficient algorithm among all considered methods for selecting the best performing subensembles. The backward elimination method has then also been applied within the proposed seasonal drought prediction system (Chapter 4). The performance measure quantified the ability of the forecasts to reproduce reference soil moisture droughts.

The variability in forecast skill among the different forecasts from the North American Multi-Model Ensemble showed differences less than 6% for the different subensembles and single model performances (Figure 4.3). On the contrary, the variability for the performance metric used for extreme precipitation and temperature indices among the different subensembles and single models based on the ENSEMBLES regional climate models ranged from 8% to 70% (Figure 2.7). Although the two metrics (i.e., Equitable Threat Score in Chapter 4 and rejection rate in Chapter 2) are very different, both of these focus on extreme conditions.

Intuitively, it is expected that realizations of different kinds of global climate models will lead to a substantially higher uncertainty than different realizations of regional climate models. The reason is that the regional climate models investigated in Chapter 2 are nested within the ERA40 reanalysis product (van der Linden and Mitchell, 2009). This implies that the boundary conditions of these models are not only the same, but they also are comparable to the observed large-scale synoptic conditions within the atmosphere. It is thus surprising that the models contained in this ensemble still create a substantial variability, which might be due to model structural differences and internal climate variability. On the other hand, the models contained in the North American Multi-Model Ensemble used in Chapter 4 are global climate models. These models are only driven by radiative forcing and are therefore able to create higher climate variability than regional climate models. However, the initialization of the global climate models becomes also an important factor when these models are used for seasonal forecasting. This initialization is comparable among the different models and dampens the variability between the forecasts of these models considerably.

Additionally, it is not surprising that the variability in forecasting seasonal soil moisture is less than the variability in reproducing long-term meteorological extremes. The reason is that the soil moisture forecasting skill is dependent on both the precise knowledge of the initial hydrologic conditions and the quality of the meteorological forecasts, which is a well accepted fact in ensemble hydrometeorological forecasting (Wood and Lettenmaier, 2008). These initial hydrologic conditions are the same among all forecasts in this work (bottom box in Figure 1.2), which further dampens the variability present in the output of global climate models. This illustrates the substantial impact of the initialization of both the global climate models and the soil moisture forecasts on the observed variability of the forecasted drought events.

5.4 Discussion of Extreme Indices

The same extreme precipitation indices investigated in Chapter 3 have been also analyzed in Chapter 2. These are the total precipitation amount above the 95^{th} percentile ($r95p$) indicating heavy precipitation events and the maximum number of consecutive dry days (CDD) as an indicator of dry spells and droughts. The latter has been termed consecutive dry time steps (CDT) in Chapter 3 because of the different temporal resolution considered in this chapter. The relative biases obtained for these two extreme indices in Chapters 3 and 2 are shown in Figure 5.3.

The relative bias of regional climate models with respect to observations amounts to more than 20% over large parts of Germany for the index $r95p$ (Figure 5.3). This is also true for the best performing model for this index, which bias is less than that of the full ensemble mean and the worst performing model. It is worth mentioning that the bias in the long-term annual precipitation amount by the best performing model is less than the bias observed for $r95p$. This implies that this model does not distribute the well-reproduced annual precipitation in a way that extremes are well captured. On the contrary, the stochastic disaggregation method introduced in Chapter 3 exhibits absolute relative bias for $r95p$ that is in general less than 5% over large parts of Germany and this threshold is only exceeded in mountainous regions such as the Black Forest and the Alps (Figure 5.3a). The absolute relative bias in these regions is also only slightly above 10% being substantially less than the bias observed for the best performing regional climate model. If the monthly values of the regional climate model with the best performance for long-term annual precipitation would have been disaggregated with the stochastic disaggregation scheme presented in Chapter 3, then the disaggregated daily values might have preserved extreme precipitation events better than the original daily values of this model. This

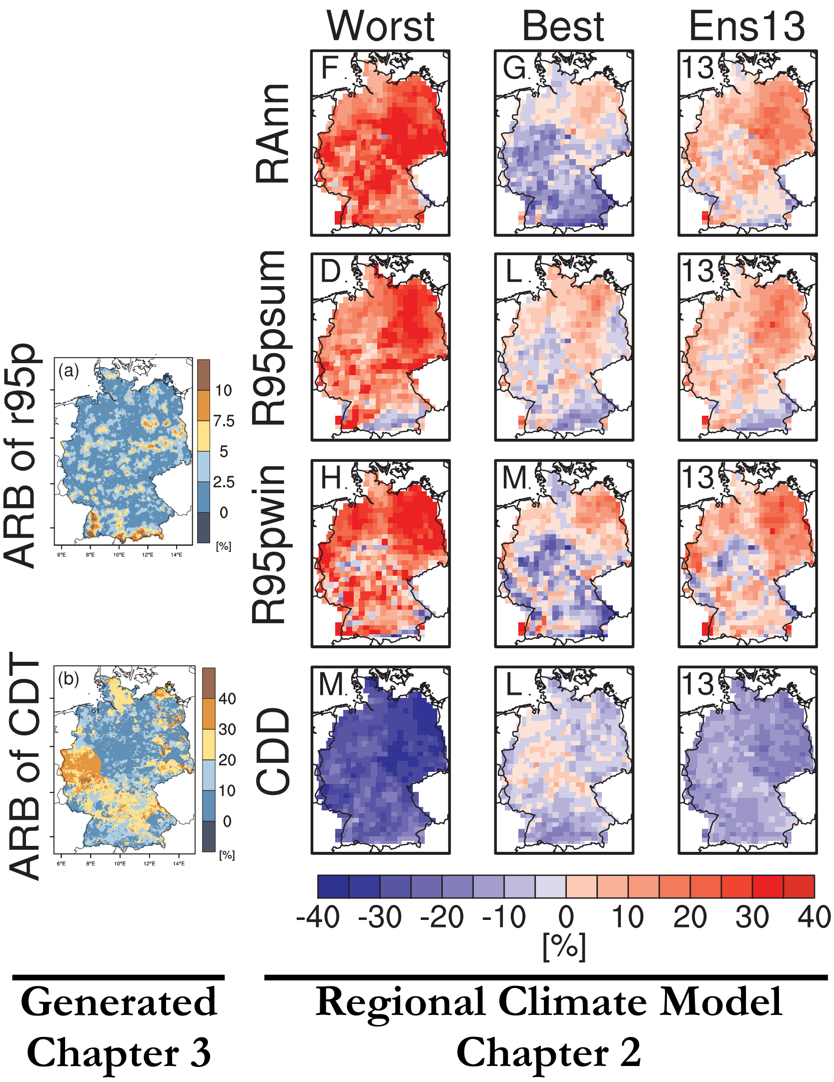


Figure 5.3: Relative biases for extreme meteorologic indices investigated in this work. The two left plots depict the absolute relative biases (ARB) for the precipitation amount above the 95th percentile ($r95p$) and the maximum number of consecutive dry time steps (CDT) of the ensemble of disaggregated monthly precipitation fields analyzed in Chapter 3 (adapted from Figure 3.11a and 3.11b). The right plots show the same indices ($r95p$ separated for summer $r95psum$ and winter $r95pwin$ and CDD which is identical to CDT for daily time steps) and the total annual precipitation $RAnn$ for the worst and best performing regional climate model and the full ensemble mean of 13 regional climate models investigated in Chapter 2 (adapted from Figure 2.4).

highlights that the stochastic disaggregation method might even be beneficial if daily climate model outputs are available.

With respect to the maximum number of consecutive dry days (CDD), the full ensemble mean of regional climate models tends to underestimate this index, but the best-performing model shows moderate positive and negative relative bias up to 10% (Figure 5.3). The absolute relative bias observed for the stochastic disaggregation method is over large parts of Germany also less than 10%, but there are also regions with a bias larger than 30% such as Western Germany (Figure 5.3b). This suggests that regional climate models can be more accurate in representing this extreme index. However, the processing steps are not identical for the two datasets. For the regional climate models investigated in Chapter 2, CDD values have been estimated for each hydrologic year and the average of the relative bias for a 40-year period is shown in Figure 5.3. On the contrary, only the maximum CDD value over the entire evaluation period has been considered in the analysis of the stochastic temporal disaggregation method in Chapter 3. In other words, the CDD values in the evaluation of the disaggregation method are based on one single value, whereas the CDD values in the evaluation of the regional climate models are based on 40 values, which is less sensitive to outliers. Moreover, the two evaluations are conducted at two different spatial resolutions. The regional climate models in Chapter 2 are analyzed at a spatial resolution of 25×25 km^2. The stochastic temporal disaggregation in Chapter 3 is conducted at a 4×4 km^2 grid. The lower resolution used in the evaluation of the regional climate models also leads to a substantially reduced variability of the reference CDD values obtained from the observations because of the larger spatial support. Overall, the differences in the sample size and spatial resolution contribute to the relatively smaller bias for the regional climate models in comparison to the bias for the stochastic disaggregation method.

In general, the analysis highlights that extreme precipitation amounts are better reproduced by stochastic realizations from the disaggregation method presented in Chapter 3 than by the realizations of state-of-the-art regional climate model simulations. For dry extremes, the best performing regional climate model showed at least the same performance as the stochastic disaggregation method. The analysis, however, was conducted in two different ways and the relatively good performance of the regional climate models could be an artifact of this.

5.5 Conclusions and Perspectives

The results and newly proposed methods in this work have broad implications for the evaluation of multi-model ensembles of climate models affecting policy recommendations and climate change impact assessments; statistical downscaling, in particular for gridded datasets with applications to large-scale hydrologic modeling; and seasonal drought forecasting.

For the first time, approaches for assessing the performance of subensembles of multi-model ensembles have been presented in this work in a unified way. The evaluation of regional climate models over Germany (taken from the ENSEMBLES project, van der Linden and Mitchell, 2009) showed that 6 out of 13 models best reproduced extreme precipitation and temperature indices estimated from the observations. Moreover, an efficient algorithm for selecting subensembles with high performances was identified. This algorithm has also been successfully employed within a seasonal drought prediction system for Europe, where the use of subensembles was also advantageous. Further multi-model ensembles that should be investigated using the proposed algorithm are those made available by the coupled model intercomparison project (Taylor K. et al., 2012), which are employed for the assessment reports of the intergovernmental panel on climate change (Flato et al., 2013).

One limitation of the evaluation of subensembles in this work is that it did not address the question why particular sets of models are performing better than others. It would be helpful for the further development of global and regional climate models to gain an understanding of the relationship between the incorporated processes in a model and its overall performance. This kind of investigation has, for example, been very successful for the development of groundwater models (Foglia et al., 2013). Achieving this mechanistic understanding of the relationship between the model structure and the model performance for climate models is, however, beyond the scope of this work because of the considerable complexity of these models and also the substantial computational resources required to run them.

The downscaling performed within this work has only focused on increasing the temporal resolution of climate model outputs but not the spatial one. Spatial downscaling methods that are based on stochastic processes such as copula-based methods (van den Berg et al., 2011) rely on a robust algorithm that generates fields of random variables with a consistent spatial covariance. The "Anchor Sampling" method introduced in Chapter 3 allows generating fields of normal distributed variates with a predefined spatial covariance at grids of any size and extent. This method has been used in this work for the temporal disaggregation of monthly precipitation, but it is in general possible to apply this method for spatial downscaling frameworks as well.

The main conclusion with respect to the seasonal drought prediction system for Europe is that the dynamic forecasts obtained from the North American Multi-Model Ensemble should be used in favor of those based on the Ensemble Streamflow Prediction approach, which has also been reported in previous studies for other regions of the world (Mo and Lettenmaier, 2014; Yuan et al., 2015). Operational drought forecasting can profit from using subensembles which only showed performance losses less than 1% in comparison to the full ensemble mean, but required only 60% of the computational demand. Drought forecasting skill was observed to be highly variable in space and time, which can be mostly attributed to the varying impact of initial hydrologic conditions. Follow-up studies should also investigate the uncertainty in the initial hydrologic conditions using methods that have been developed for purely observation-based datasets such as the Reverse Ensemble Streamflow Prediction approach (Wood and Lettenmaier, 2008; Shukla and Lettenmaier, 2011; Shukla et al., 2013).

In summary, this work tried to advance the understanding of how to increase the efficiency of multi-model ensembles for hydrologic impact assessment at large spatial scales such as seasonal drought prediction over Europe. It also added to the stochastic modeling of precipitation on large spatial grids with an application to the temporal disaggregation of monthly precipitation to daily values. The potential benefits of the presented methods were only shown for a limited number of examples. Follow-up studies should apply the presented methods, which are general, to other datasets and also use them for other applications such as, for instance, spatial downscaling of climate model outputs.

Appendices

A Anchor sampling

A crucial step during the generation procedure is the correct spatial sampling of a multivariate normal distribution required in Generation step 2.) (Figure 3.2).

The standard approach would be to: A) generate a vector $\overline{\xi}$ of i.i.d. standard normal values and, B) apply an affine transformation

$$\xi = \mu + \mathbf{C}\overline{\xi}, \tag{A.1}$$

where μ is a vector of mean values and $\mathbf{C}$ is the lower Cholesky factor of a cross-covariance matrix $\mathcal{B}$. The resulting vector ξ has the following multivariate normal distribution:

$$\xi \sim \mathcal{N}(\mu, \mathbf{C}\mathbf{C}^{\mathsf{T}}) = \mathcal{N}(\mu, \mathcal{B}). \tag{A.2}$$

Advantages of this traditional approach are that the covariances among all cells are considered and that the Cholesky factor has to be calculated only once. $\mathcal{B}$ is a $N \times N$ matrix (N is the number of cells) and may contain a huge number of entries for large grids as in this study. For large matrices, there is no guarantee that the Cholesky factor $\mathbf{C}$ can be computed because the estimated cross-covariance matrices $\mathcal{B}$ from the observations might not be positive definite.

An alternative approach to this standard method is to apply a conditional or sequential Gaussian sampling where the entries of ξ are generated sequentially one after the other. Sequential Gaussian sampling is discussed in detail by Dimitrakopoulos and Luo (2004) and will be shortly summarized here.

Let $\xi(i),\ i = 1, \ldots, N$ be a vector to be sampled from a multivariate normal distribution. In this study, i represents an indexing of locations, i.e., grid cell. Let δ be a permutation of these indices which defines a path through the domain assigning the order in which the cells are going to be generated. Let $\Lambda_{i-1} = \{\delta(j) |\ j = 1, \ldots, i-1\}$ denote the set of cells that were generated previously to cell i. Using this notation, the i^{th} cell can be generated as follows (equation 4 in Dimitrakopoulos and Luo (2004))

$$\xi(i|\Lambda_{i-1}) = E\{\xi(i)|\Lambda_{i-1}\} + \sqrt{Var\{\xi(i|\Lambda_{i-1})\}} \cdot \overline{\xi}(i), \tag{A.3}$$

where $E\{\xi(i)|\Lambda_{i-1}\}$ and $Var\{\xi(i|\Lambda_{i-1})\}$ are the mean and variance, respectively, conditioned on the set Λ_{i-1} of already generated cells. Dimitrakopoulos and Luo (2004) furthermore noted that the conditional mean and variance can be calculated by

$$E\{\xi(i)|\Lambda_{i-1}\} = m_i + \mathcal{B}_{i\Lambda_{i-1}}\mathcal{B}^{-1}_{\Lambda_{i-1}\Lambda_{i-1}}(\xi_{\Lambda_{i-1}} - m_{\Lambda_{i-1}}), \tag{A.4}$$

$$Var\{\xi(i|\Lambda_{i-1})\} = \mathcal{B}_{ii} - \mathcal{B}_{i\Lambda_{i-1}}\mathcal{B}^{-1}_{\Lambda_{i-1}\Lambda_{i-1}}\mathcal{B}_{\Lambda_{i-1}i}, \tag{A.5}$$

where m_i and $\mathcal{B}_{ii}$ are the mean and variances of cell i, respectively. In this study, m_i and $m_{\Lambda_{i-1}}$ are equal to zero because a normal distribution with zero mean is assumed in equation 3.8. $\xi_{\Lambda_{i-1}}$ is the vector of already generated values. $\mathcal{B}_{i\Lambda_{i-1}}$ is the covariance of the current cell i with the already generated cells. $\mathcal{B}_{\Lambda_{i-1}\Lambda_{i-1}}$ is the cross-covariance matrix of the already generated cells.

Dimitrakopoulos and Luo (2004) further discussed algorithms to reduce the computational cost of the sequential Gaussian sampling given theoretical variance-covariance matrices. In the following paragraphs, a modification of the sequential Gaussian sampling, called “Anchor Sampling”, is introduced and employed in this study. This modification is different from the algorithms discussed in Dimitrakopoulos and Luo (2004) and provides reasonable results in a “real-world” application.

The numerical demanding step in the computation of equations A.4 and A.5 is the calculation of the inverse $\mathcal{B}^{-1}_{\Lambda_{i-1}\Lambda_{i-1}}$. This step is not guaranteed to be computable, if the set Λ_{i-1} becomes large. In this study, that occurred when the size of the set Λ_{i-1} was exceeding 500. To restrict the size $L = |\Lambda_{i-1}|$, only cells in the neighborhood of the cell i are considered, i.e., cells that are closer than a given threshold D. As a rule of thumb, L should not be less than 100 because the cells would be too loosely connected to resemble the covariance structure correctly. Furthermore, L should not be greater than 400 because this would lead to high computational costs. A value of 200 using one covariance matrix for all months and 150 using two covariance matrices (one for each summer and winter season) has proven to be a good compromise between computational efficiency and reliable representation of the covariance structure (see supplementary material for benchmark).

This local conditioning leads to a potential underrepresentation of correlations with distant cells as well as global characteristics. A set Ω of randomly distributed anchor cells is introduced to circumvent this drawback and to ensure a globally consistent pattern. This anchor set is small (< 100) compared to the number of cells in Λ_{i-1}. The standard normal variates at these anchor cells are generated first employing the standard approach, i.e., using only information of the anchoring cells (equation A.1). The anchor cells are then removed from the path δ through the domain. The remaining cells are generated like before, except that the set of anchor cells is always added to the set of neighboring cells Λ_{i-1}. Hence, the new set $\widetilde{\Lambda}_{i-1}$ the cell i is conditioned on is

$$\widetilde{\Lambda}_{i-1} = \Lambda_{i-1} \cup \Omega. \tag{A.6}$$

The number of anchor cells used in this study is 20 (see supplementary material for benchmark).

The neighborhood threshold D is typically selected such that the total number of cells in the neighborhood is greater than L. If there are more than L cells already generated in the neighborhood, a subset of these cells has to be selected such that $\tilde{\Lambda}_{i-1}$ contains only L cells. The probability of selecting cell j as neighboring cell of cell i is

$$\mathbf{P}_i(j) \propto \frac{1}{d_{ij}}. \tag{A.7}$$

In this study, d_{ij} is the distance of the already generated cell i to cell j in km. The threshold D used in this study is 200 km using one covariance matrix for all months and 175 km using two matrices for different seasons, which has been proven to yield reliable results (see supplementary material for benchmark). It is worth noting that this approach is able to correctly reproduce the correlation structure with cells that are separated by distance of more than 200 km and 175 km. For example, the correlation coefficients ρ for cells separated by distances of 300 km are comparable for the observations and simulations at the weekly scale (see Section 3.5.5).

B Equations for ARB and MARB

The absolute relative bias (ARB) discussed in Sections 3.5.4 are calculated as follows

$$x = \frac{|m - o|}{o}. \tag{B.1}$$

Here, m and o denote statistics obtained from modeled and observed precipitation, respectively. If multiple of such statistics are available, like for example in the case of deciles $D_1, \ldots, D_9$ (Section 3.5.2), the ARB of these are summarized as arithmetic mean (hence MARB) as follows

$$x = \frac{1}{N} \sum_{i=1}^{N} \frac{|m_i - o_i|}{o_i}. \tag{B.2}$$

Here, i is an index of different statistics obtained from modeled and observed precipitation.

C Technical Material for "Stochastic Temporal Disaggregation of Monthly Precipitation for Regional Gridded Data Sets"

C.1 Algorithm for the Estimation of the Intensity Classes

One step in the proposed method is the estimation of the intensity classes (Figure 3.2: Estimation, step 3). The thresholds for these intensity classes are estimated to be invariant in space and are selected such that the following criteria are fulfilled. First, each class should contain at least 5% of all data points. Second, each class should contain on average at least 150 data points per cell to guarantee a robust estimation of the cdfs of the weights. The exact number of data points per cell in each class can deviate from this criteria because of the spatial variability of precipitation.

The following algorithm was used to find intensity class thresholds in such a way that the average number of data points per class is fulfilling the aforementioned criteria. The required variables of this algorithm are (variable names in parentheses): the lower threshold of the current intensity class (l), the upper threshold of the current intensity class (u), the average number of data points between l and u over all cells (a), the target number of data points between l and u per cell (t), the number of classes (c), and the maximum value of the given precipitation field (m).

The detailed steps of the algorithm are as follows:

1. Initialize all variables. Let t equal to the maximum of 150 and 5% of the number of given time steps; a and l equal to zero; and u and c equal to one. Store l as the lower intensity threshold for the first class.
2. Increase u by 10% and re-estimate a.
3. Repeat step 2 until a is greater than t or until u is greater than m.

4a. If u is less than m, apply a nested loop algorithm to adapt u until a is equal to t. Store u as the upper intensity class threshold for class c and the lower intensity threshold for the next class $c + 1$. Increment c by one. Set l to u and go to step 2.

4b. If u is greater than m, store m as the upper intensity threshold for the last class. All intensity class thresholds have been found. The final number of classes determined is c.

C.2 Parameter Estimation of Spatial Sampling

The spatial sampling algorithm described in Appendix A depends on three parameters. These are the number of anchor cells $A = |\Omega|$, the distance threshold of the neighborhood D, and the number of neighborhood cells L. The following benchmark was performed using the covariance matrix disaggregating two-day precipitation estimated for all calendar months to identify a well suitable setting of these parameters.

The Frobenius norm for a matrix $\mathcal{A} = (a_{ij})$ was used to quantify the difference between the observed and simulated covariance matrices. It is defined as

$$\|\mathcal{A}\| := \sqrt{\sum_{i,j=1}^{N} |a_{ij}|^2}, \tag{C.1}$$

where N is the total number of grid cells. This norm was calculated for the difference matrix $\mathcal{A}_D$

$$\mathcal{A}_D = \mathcal{A}_S - \mathcal{A}_O, \tag{C.2}$$

where $\mathcal{A}_S$ and $\mathcal{A}_O$ denote the cross-covariance matrices of the whole field for the simulations and observations, respectively. The matrix $\mathcal{A}_S$ depends on the parameters A, L, and D. $\mathcal{A}_S$ would be the identity matrix, if A and L would be set to zero because the cells would be sampled independently of each other. On the contrary, $\mathcal{A}_S$ would be very close to $\mathcal{A}_O$, if A would be set to the total number of cells. In this case, the difference would only occur from the intrinsic noise of the random sampling. The proposed method, however, would become a standard approach in such a setting and thus would be numerically unstable.

The Frobenius norm was estimated for 45 different parameter settings. These are all combinations of

$$\begin{aligned} A &\in \{0, 20, 40, 60, 80\}, \\ D &\in \{100, 200, 300\},\ [\text{km}] \text{ and} \\ L &\in \{100, 200, 300\}. \end{aligned}$$

An ensemble of ten different paths δ, i.e., ten different realizations of the sequential sampling (see Appendix A), were determined for each parameter setting. The minimum, median, and maximum Frobenius norm for each parameter setting were determined (Figure C.1).

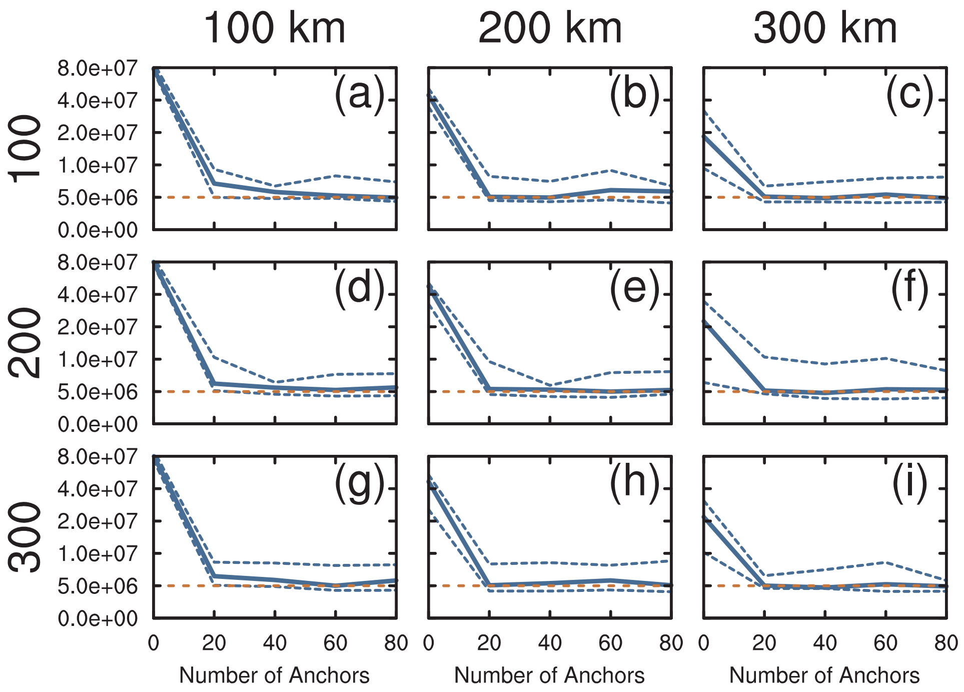


Figure C.1: Frobenius norm of the difference matrix between observed and simulated cross-covariance matrix for different parameter settings. The distance thresholds D are 100 km, 200 km, and 300 km for the first, second, and third column, respectively. The number of neighboring cells L are 100, 200, and 300 for the first, second, and third row, respectively. The selected number of anchor cells A are 0, 20, 40, 60, and 80. These are displayed on the x-Axis of each plot. The 5.0×10^8 line (red dashed line) is displayed to ease the comparison among different plots.

The number of anchor cells A has a similar impact on the Frobenius Norm for each combination of D and L. The Frobenius norm decreases substantially when A is set from 0 to 20 but stays constant when A is further increased. This implies that 20 anchor cells are sufficient for the spatial extent of Germany and this value is also chosen in the present study. In general, the impact of A does not strongly depend on the spatial resolution of the study area but more on its orographic features and extent. A heterogeneous mountainous study domain will require more anchor cells than a homogeneous flat one.

The distance threshold D has a high influence on the Frobenius norm when no anchor cells are considered. For example, the Frobenius norm decreases from 8.0×10^7 to 2.0×10^7 when the distance threshold is increased from 100 km to 300 km (Figure C.1: left and right column). This result could have been expected because a larger distance threshold implies that the connectivity of the cells is increased. This effect, however, vanishes when A is larger than zero. This highlights the importance of anchor cells because they ensure a connectivity among the cells even if the distance threshold D is relatively small. A distance threshold of 200 km is selected in the present study because there is a slight decrease in the Frobenius norm between a D of 100 km and 200 km for A of 20. Additionally, choosing 300 km would imply that almost all cells including anchor cells are in the neighborhood of grid cells located in Central Germany. This represents an unrealistic testbed for the sequential sampling with regard to spatial data sets of even larger spatial extent (e.g., data sets for continental United States or Europe) because the distance threshold would have to be set to at least 1000 km such that all cells are also neighborhood cells for data sets of this large extent.

The impact of the number of neighboring cells L seems to be negligible as the Frobenius norm has similar values for each setting of D and A (Figure C.1: compare rows). However, the parameter L has the highest impact on the computational costs of the proposed method. This stems from the fact that the computational cost of the inversion of the matrices in equations A.4 and A.5 scale quadratically with L. A number of 200 neighboring cells has proven to yield good results in the present study at reasonable computational costs. This choice is made because it leads to a good balance between the number of selected neighboring cells L and the total number of cells within the distance threshold D. The ratio between these two is around 2.5% in the present study. If this ratio is too high, the algorithm is inefficient because of the increase in computational expenses. If this ratio is too low, the connectivity in the neighborhood decreases and too much random noise will occur.

As a result of this sensitivity analysis, the selected parameter in this study are a D of 200 km, A of 20, and L of 200 for all months. The sample size available for estimating the covariance matrices during summer and winter months is only half of that used for the estimate for all months. For this reason, the estimate is poorer than the one taking all data points into account and the parameter L was changed to 150 to accommodate for this fact. As a consequence, a D of 175 km was used to maintain the ratio of 2.5% outlined above.

C.3 Spatial Clustering of CDFs

In the proposed method, the cdfs $\mathcal{F}$ of the weights between two scales are estimated for each location i and intensity class c (equation 3.3a). Hence, the number of known cdfs in the proposed method is $N \times C$, where N is the number of locations and C the number of intensity classes over all scales. This is a huge number because of the large number of grid cells, i.e., $N \gg 20,000$. The next steps in the development of this method is the parametrization of the cdfs $\mathcal{F}$ which will be difficult to obtain with this large number of unknowns. A reasonable regionalization of these functions must be found to achieve a parsimonious parametrization of the method. It has to be pointed out that this drawback is not present in the proposed method because of the usage of empirical cdfs.

A relationship between specific location characteristics and the cdfs $\mathcal{F}$ have to be found to derive a regionalization of these functions. A k-means clustering Wilks (2011) was performed on the precipitation distribution functions and the cdfs $\mathcal{F}$ of one scale and intensity class (Figure C.2) to test whether such a relationship could be established. The drawback of this clustering approach is that the number of clusters has to be predefined. The advantage in turn is that it is non-hierarchical, i.e., cells can be moved from one cluster to another. In this study, four clusters have been chosen beforehand because this number of clusters captures the main spatial patterns for both the precipitation cdfs and the cdfs of the weights.

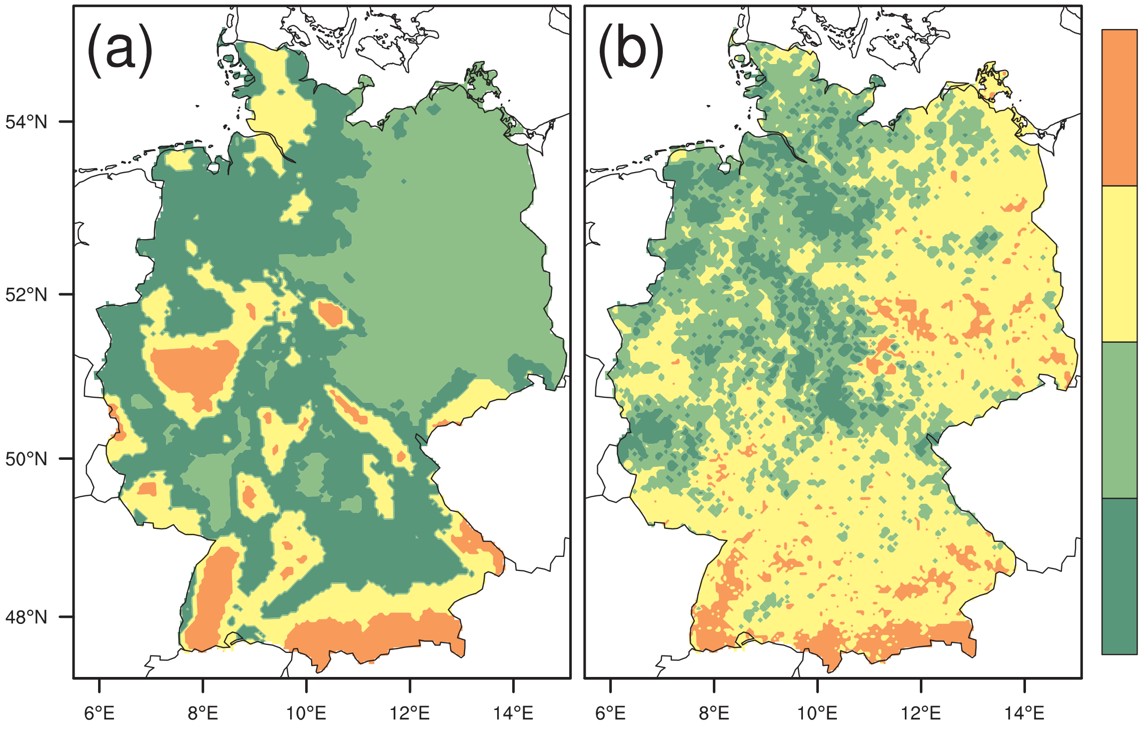


Figure C.2: Spatial distribution of k-means cluster with $k = 4$ of (a) precipitation distribution functions and (b) cdfs $\mathcal{F}$ between two day and daily scale for one intensity class over Germany.

The clusters derived for the precipitation distribution functions can be associated with specific regions (Figure C.2a). For example, the precipitation cdfs in mountainous regions in Central and South Germany, the lowlands in North-East Germany, and the lowlands in North-West Germany are grouped in separate clusters (cluster 4, 2, and 1, respectively). Overall, the general spatial pattern is similar to that of long-term annual precipitation (see Figure 3.3). Both of these patterns show a strong dependence on elevation. This relationship on elevation can be exploited to identify a regionalization for the cdfs of precipitation. Using this regionalization for the cdfs $\mathcal{F}$ would imply that one assumes that the cdfs of the weights cluster in the same way as the ones of precipitation.

This assumption is tested for the cdfs $\mathcal{F}$ between the two day and daily scale for one intensity class. The clusters of these cdfs can also be associated with specific regions (Figure C.2b). For example, the cdfs grouped in cluster 4 are mainly located in the Black Forest mountain range in South-West Germany and the Prealps in Southern Germany. The cdfs attributed to cluster 1 and 2 are also mainly located in the lowlands in North-West Germany, whereas cdfs in East Germany are mostly grouped in cluster 3. The general pattern, however, is very different from that obtained for the precipitation cdfs (compare Figure C.2a and C.2b). This implies that there might be other variables than only elevation influencing the shape of the cdfs $\mathcal{F}$, such as for example the probability of dry days (see Figure 3.10a).

Further research is required to implement and test a parsimonious parametrization of the proposed method. Any regionalized parameter will lead to higher bias between generated and observed precipitation cdfs because assumptions have to be formulated that might not hold in each case. This study can be regarded as a benchmark for methods with regionalized parameters because the full capability of the proposed method by employing empirical distribution functions is investigated here.

D Equitable Threat Score

Forecast verification for discrete events (e.g., a drought event) is commonly carried out using measures that are based on a 2×2 contingency table (Wilks, 2011). In this study, we use the Equitable Threat Score (ETS) as skill measure, which is defined as

$$ETS = 100\frac{a - a_{ref}}{a - a_{ref} + b + c}, \tag{D.1}$$

where a is the number of drought events that occur in both the forecast and the reference dataset (commonly called hits), b is the number of drought events that occur in the forecast but not in the reference dataset (commonly called false alarms), and c is the number of droughts that occur not in the forecast but in the reference dataset (commonly called misses). a_{ref} is defined as

$$a_{ref} = \frac{(a + b)(a + c)}{n}, \tag{D.2}$$

where n is the total number of time steps. ETS is used in this study because it condenses the hit rate $(a/(a + c))$ and the false alarm rate $(b/(a + b))$ into one metric. An ETS of 100% indicates a hit rate of 1 and a false alarm rate of 0, which means that all drought events are forecasted perfectly.

Appendix D:
Equitable Threat Score

E Drought severity and area

Two drought characteristics are evaluated during the drought development and recovery phase. These are the fraction of correctly forecasted drought area and the drought severity of this area. For a given time step t, the former is defined as

$$A(t) = \frac{a(t)}{a(t) + c(t)}, \tag{E.1}$$

where $a(t)$ is the number of grid cells under drought both in the forecast and the reference dataset at time step t and $c(t)$ is the number of grid cells under drought that occur not in the forecast but in the reference dataset at time step t. It is worth mentioning that this area is equivalent to the hit rate estimated over space.

The drought severity is calculated for the grid cells that exhibit a drought both in the forecast and the reference dataset. For a given time step t, the drought severity is defined as

$$S(t) = \sum_{i \in a(t)} [\tau - \mathrm{SMI}_i(t)]_+, \tag{E.2}$$

where τ is the SMI drought threshold (here 0.2), $(\cdot)_+$ is the positive part function, and $a(t)$ is defined as above. A large deviation from the drought threshold leads to higher severity indicating a more severe drought. The severity of the forecast is then normalized by that of the reference dataset to make them comparable among different drought events.

Bibliography

Andreadis, K. M., Clark, E. A., Wood, A. W., Hamlet, A. F., and Lettenmaier, D. P. Twentieth-Century Drought in the Conterminous United States. *Journal of Hydrometeorology*, 6(6):985–1001, 2005. ISSN 1525-755X. doi: 10.1175/JHM450.1.

Bárdossy, A. and Pegram, G. G. S. Copula based multisite model for daily precipitation simulation. *Hydrology and Earth System Sciences*, 13(12):2299–2314, 2009. doi: 10.5194/hess-13-2299-2009.

Beniston, M., Stephenson, D., Christensen, O., Ferro, C., Frei, C., Goyette, S., Halsnaes, K., Holt, T., Jylhä, K., Koffi, B., Palutikof, J., Schöll, R., Semmler, T., and Woth, K. Future extreme events in European climate: an exploration of regional climate model projections. *Climatic Change*, 81 (1):71–95, 2007. ISSN 0165-0009. doi: 10.1007/s10584-006-9226-z.

Blöschl, G., Reszler, C., and Komma, J. A spatially distributed flash flood forecasting model. *Environmental Modelling & Software*, 23(4):464–478, 2008. ISSN 1364-8152. doi: 10.1016/j.envsoft.2007.06.010.

Boberg, F., Berg, P., Thejll, P., Gutowski, W., and Christensen, J. Improved confidence in climate change projections of precipitation further evaluated using daily statistics from ENSEMBLES models. *Climate Dynamics*, 35:1509–1520, 2010. ISSN 0930-7575. doi: 10.1007/s00382-009-0683-8.

Bonan, G. Ecological Climatology: Concepts and Applications. Cambridge University Press, 2008. ISBN 9780521872218.

Booij, M. Impact of climate change on river flooding assessed with different spatial model resolutions. *Journal of Hydrology*, 303(1–4):176–198, 2005. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2004.07.013.

Bormann, H., Ahlhorn, F., and Klenke, T. Adaptation of water management to regional climate change in a coastal region – Hydrological change vs. community perception and strategies. *Journal of Hydrology*, 454–455(0):64–75, 2012. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2012.05.063.

Bougeault, P., Toth Zoltan, Bishop Craig, Brown Barbara, Burridge David, Chen De Hui, Ebert Beth, Fuentes Manuel, Hamill Thomas M., Mylne Ken, Nicolau Jean, Paccagnella Tiziana, Park Young-Youn, Parsons David, Raoult Baudouin, Schuster Doug, Dias Pedro Silva, Swinbank

Richard, Takeuchi Yoshiaki, Tennant Warren, Wilson Laurence, and Worley Steve. The THORPEX Interactive Grand Global Ensemble. *Bulletin of the American Meteorological Society*, 91(8): 1059–1072, 2010. ISSN 0003-0007. doi: 10.1175/2010BAMS2853.1.

Buizza, R. The tigge global, medium-range ensembles. Technical report, ECMWF, 2015. URL `http://old.ecmwf.int/publications/library/ecpublications/_pdf/tm/701-800/tm739.pdf`. *ECMWF Technical Memorandum*.

Burton, A., Fowler, H. J., Kilsby, C. G., and O'Connell, P. E. A stochastic model for the spatial-temporal simulation of nonhomogeneous rainfall occurrence and amounts. *Water Resources Research*, 46(11), 2010. ISSN 1944-7973. doi: 10.1029/2009WR008884.

Chandler, R. E. Exploiting strength, discounting weakness: combining information from multiple climate simulators. *Philosophical Transactions of the Royal Society A*, 371(1991), May 2013. doi: 10.1098/rsta.2012.0388.

Charles, S. P., Bates, B. C., Whetton, P. H., and Hughes, J. P. Validation of downscaling models for changed climate conditions: case study of southwestern Australia. *Climate Research*, 12:1–14, 1999.

Chow, V., Maidment, D., and Mays, L. Applied Hydrology. McGraw-Hill series in water resources and environmental engineering. Tata McGraw-Hill Education, 1988. ISBN 9780070702424.

Christensen, O., Drews, M., Christensen, J., Dethloff, K., Ketelsen, K., Hebestadt, I., and Rinke, A. The HIRHAM regional climate model version 5(b). Technical Report 06-17, Dan. Meteorol. Inst., 2006.

Clark, M. P. and Slater, A. G. Probabilistic Quantitative Precipitation Estimation in Complex Terrain. *Journal of Hydrometeorology*, 7(1):3–22, Feb 2006. ISSN 1525-755X. doi: 10.1175/JHM474.1.

Collins, M., Booth, B. B. B., Harris, G. R., Murphy, J. M., Sexton, D. M. H., and Webb, M. J. Towards quantifying uncertainty in transient climate change. *Climate Dynamics*, 27(2-3):127–147, 2006. doi: 10.1007/s00382-006-0121-0.

COPA-COGECA. Assessment of the Impact of the Heat Wave and Drought of the Summer 2003 on Agriculture and Forestry, 2003. In Committee of Agricultural Organisations in the European Union General Committee for Agricultural Cooperation in the European Union, Brussels, 15 pp.

Davison, A. C. and Smith, R. L. Models for Exceedances over High Thresholds. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52(3):393–442, 1990.

Day, G. Extended Streamflow Forecasting Using NWSRFS. *Journal of Water Resources Planning and Management*, 111(2):157–170, 1985. doi: 10.1061/(ASCE)0733-9496(1985)111:2(157).

Deidda, R., Badas, M. G., and Piga, E. Space–time multifractality of remotely sensed rainfall fields. *Journal of Hydrology*, 322(1–4):2–13, 2006. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2005.02.036.

Deser, C., Phillips, A., Bourdette, V., and Teng, H. Uncertainty in climate change projections: the role of internal variability. *Climate Dynamics*, 38(3-4):527–546, 2012. ISSN 0930-7575. doi: 10.1007/s00382-010-0977-x.

Dimitrakopoulos, R. and Luo, X. Generalized Sequential Gaussian Simulation on Group Size v and Screen-Effect Approximations for Large Field Simulations. *Mathematical Geology*, 36(5):567–591, July 2004.

Dirmeyer, P. A., Guo, Z., and Gao, X. Comparison, validation, and transferability of eight multiyear global soil wetness products. *Journal of Hydrometeorology*, 5(6):1011–1033, December 2004. ISSN 1525-755X. doi: 10.1175/JHM-388.1.

Doblas Reyes, F. J., Hagedorn, R., and Palmer, T. N. The rationale behind the success of multi-model ensembles in seasonal forecasting–II. Calibration and combination. *Tellus A*, 57(3):234–252, 2005.

Döll, P., Kaspar, F., and Lehner, B. A global hydrological model for deriving water availability indicators: model tuning and validation. *Journal of Hydrology*, 270(1-2):105 – 134, 2003. ISSN 0022-1694. doi: http://dx.doi.org/10.1016/S0022-1694(02)00283-4.

D'Onofrio D., Palazzi E., von Hardenberg J., Provenzale A., and Calmanti S. Stochastic rainfall downscaling of climate models. *Journal of Hydrometeorology*, 2014. ISSN 1525-755X. doi: 10.1175/JHM-D-13-096.1.

Dutra, E., Magnusson, L., Wetterhall, F., Cloke, H. L., Balsamo, G., Boussetta, S., and Pappenberger, F. The 2010–2011 drought in the Horn of Africa in ECMWF reanalysis and seasonal forecast products. *International Journal of Climatology*, 33(7):1720–1729, 2013. ISSN 1097-0088. doi: 10.1002/joc.3545.

Ebtehaj, M. and Foufoula-Georgiou, E. Orographic signature on multiscale statistics of extreme rainfall: A storm-scale study. *Journal of Geophysical Research: Atmospheres*, 115(D23):n/a–n/a, 2010. ISSN 2156-2202. doi: 10.1029/2010JD014093.

Farda, A., Déué, M., Somot, S., Horányi, A., Spiridonov, V., and Tóth, H. Model ALADIN as regional climate model for Central and Eastern Europe. *Studia Geophysica et Geodaetica*, 54:313–332, 2010. ISSN 0039-3169. doi: 10.1007/s11200-010-0017-7.

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S., Collins, W., Cox, P., Driouech, F., Emori, S., Eyring, V., Forest, C., Gleckler, P., Guilyardi, E., Jakob, C., Kattsov, V., Reason, C., and Rummukainen, M. Evaluation of climate models. In Stocker, T., Qin, D., Plattner, G.-K., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P., editors, *In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assess- ment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

Foglia, L., Mehl, S. W., Hill, M. C., and Burlando, P. Evaluating model structure adequacy: The case of the maggia valley groundwater system, southern switzerland. *Water Resources Research*, 49(1):260–282, 2013. ISSN 1944-7973. doi: 10.1029/2011WR011779.

Fowler, H., Cooley, D., Sain, S., and Thurston, M. Detecting change in UK extreme precipitation using results from the climateprediction.net BBC climate change experiment. *Extremes*, 13(2):241–267, 2010. ISSN 1386-1999. doi: 10.1007/s10687-010-0101-y.

Frei, C., Schöll, R., Fukutome, S., Schmidli, J., and Vidale, P. L. Future change of precipitation extremes in Europe: Intercomparison of scenarios from regional climate models. *Journal of Geophysical Research: Atmospheres*, 111(D6):n/a–n/a, 2006. ISSN 2156-2202. doi: 10.1029/2005JD005965.

Giorgi, F. and Mearns, L. O. Calculation of Average, Uncertainty Range, and Reliability of Regional Climate Changes from AOGCM Simulations via the Reliability Ensemble Averaging (REA) Method. *Journal of Climate*, 15:1141–1158, May 2002. ISSN 0894-8755. doi: 10.1175/1520-0442(2002)015.

Güntner, A., Olsson, J., Calver, A., and Gannon, B. Cascade-based disaggregation of continuous rainfall time series: the influence of climate. *Hydrology and Earth System Sciences*, 5(2):145–164, 2001. doi: 10.5194/hess-5-145-2001.

Haberlandt, U., Hundecha, Y., Pahlow, M., and Schumann, A. Rainfall generators for application in flood studies. In Schumann, A., editor, *Flood Risk Assessment and Management*, pages 117–147. Springer Netherlands, 2011.

Hargreaves, G. H. and Samani, Z. A. Reference crop evapotranspiration from temperature. *Applied Engineering in Agriculture*, 1(2):96–99, 1985. doi: 10.13031/2013.26773.

Harris, I., Jones, P., Osborn, T., and Lister, D. Updated high-resolution grids of monthly climatic observations – the CRU TS3.10 Dataset. *International Journal of Climatology*, 34(3):623–642, 2014. ISSN 1097-0088. doi: 10.1002/joc.3711.

Haugen, J. and Haakenstad, H. Validation of HIRHAM version 2 with 50km and 25km resolution. Technical Report 9, Norw. Meteorol. Inst., 2005.

Hay, L. E., Clark, M. P., Wilby, R. L., Gutowski, W. J., Leavesley, G. H., Pan, Z., Arritt, R. W., and Takle, E. S. Use of Regional Climate Model Output for Hydrologic Simulations. *Journal of Hydrometeorology*, 3(5):571–590, Oct 2002. ISSN 1525-755X. doi: 10.1175/1525-7541(2002)003.

Hay, L. E., Wilby, R. L., and Leavesley, G. H. A comparison of Delta Change and Downscaled GCM Scenarios for three mountainous basins in the United States. *JAWRA Journal of the American Water Resources Association*, 36(2):387–397, 2000. ISSN 1752-1688. doi: 10.1111/j.1752-1688.2000.tb04276.x.

Haylock, M. R., Hofstra, N., Klein Tank, A. M. G., Klok, E. J., Jones, P. D., and New, M. A european daily high-resolution gridded data set of surface temperature and precipitation for 1950-2006. *Journal of Geophysical Research: Atmospheres*, 113(D20), 2008. ISSN 2156-2202. doi: 10.1029/2008JD010201.

Herrera, S., Fita, L., Fernández, J., and Gutiérrez, J. M. Evaluation of the mean and extreme precipitation regimes from the ENSEMBLES regional climate multimodel simulations over Spain. *Journal of Geophysical Research*, 115(D21117):1–13, 2010.

Hrachowitz, M., Savenije, H., Blöschl, G., McDonnell, J., Sivapalan, M., Pomeroy, J., Arheimer, B., Blume, T., Clark, M., Ehret, U., Fenicia, F., Freer, J., Gelfan, A., Gupta, H., Hughes, D., Hut, R., Montanari, A., Pande, S., Tetzlaff, D., Troch, P., Uhlenbrook, S., Wagener, T., Winsemius, H., Woods, R., Zehe, E., and Cudennec, C. A decade of predictions in ungauged basins (pub) - a review. *Hydrological Sciences Journal*, 58(6):1198–1255, 2013. doi: 10.1080/02626667.2013.803183.

Hundecha, Y. and Bárdossy, A. Statistical downscaling of extremes of daily precipitation and temperature and construction of their future scenarios. *International Journal of Climatology*, 28(5): 589–610, 2008. ISSN 1097-0088. doi: 10.1002/joc.1563.

Hundecha, Y., Pahlow, M., and Schumann, A. Modeling of daily precipitation at multiple locations using a mixture of distributions to characterize the extremes. *Water Resources Research*, 45(12), 2009. ISSN 1944-7973. doi: 10.1029/2008WR007453.

Jacob, D., Van Den Hurk, B. J. J. M., Andræ, U., Elgered, G., Fortelius, C., Graham, L. P., et al. A comprehensive model inter-comparison study investigating the water budget during the BALTEX-PIDCAP period. *Meteorology and Atmospheric Physics*, 77:19–43, 2001. doi: 10.1007/s007030170015.

Jaeger, E. B., Anders, I., Lüthi, D., Rockel, B., Schär, C., and Seneviratne, S. I. Analysis of ERA40-driven CLM simulations for Europe. *Meteorologische Zeitschrift*, 17(4):349–367, 2008. doi: 10.1127/0941-2948.

Jones, P. W. First- and Second-Order Conservative Remapping Schemes for Grids in Spherical Coordinates. *Monthly Weather Review*, 127(9):2204–2210, Sep 1999. ISSN 0027-0644. doi: 10.1175/1520-0493(1999)127.

Jun, M., Knutti, R., and Nychka, D. W. Spatial Analysis to Quantify Numerical Model Bias and Dependence. *Journal of the American Statistical Association*, 103(483):934–947, 2008. doi: 10.1198/016214507000001265.

Kirtman, B. P., Power, S., Adedoyin, J., Boer, G., Bojariu, R., Camilloni, I., Doblas-Reyes, F., Fiore, A., Kimoto, M., Meehl, G., Prather, M., Sarr, A., Schär, C., Sutton, R., van Oldenbourgh, G., Vecchi, G., and Wang, H. Near-term Climate Change: Projections and Predictability. In Stocker, T., Qin, D., Plattner, G.-K., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P., editors, *Climate Change 2013: The Physical Science Basis. Contribution of Working Group 1 to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, 2013.

Kirtman, B. P., Min, D., Infanti, J. M., Kinter, J. L., Paolino, D. A., Zhang, Q., van den Dool, H., Saha, S., Mendez, M. P., Becker, E., Peng, P., Tripp, P., Huang, J., DeWitt, D. G., Tippett, M. K., Barnston, A. G., Li, S., Rosati, A., Schubert, S. D., Rienecker, M., Suarez, M., Li, Z. E., Marshak, J., Lim, Y.-K., Tribbia, J., Pegion, K., Merryfield, W. J., Denis, B., and Wood, E. F. The North American Multi-Model Ensemble (NMME): Phase-1 Seasonal to Interannual Prediction, Phase-2 Toward Developing Intra-Seasonal Prediction. *Bulletin of the American Meteorological Society*, Aug 2014. ISSN 0003-0007. doi: 10.1175/BAMS-D-12-00050.1.

Kjellström, E., Bärring, L., Gollvik, S., Hansson, U., Jones, C., Samuelsson, P., Rummukainen, M., Ullerstig, A., U., W., and Wyser, K. A 140-year simulation of European climate with the new version of the Rossby Centre regional atmospheric climate model (RCA3). Technical Report 108, SMHI, SE-60176 Norrköping, Sweden, 2005.

Knutti, R., Abramowitz, G., Collins, M., Eyring, V., Gleckler, P. J., Hewitson, B., and Mearns, L. Good Practice Guidance Paper on Assessing and Combining Multi Model Climate Projections , 2010. Rep. of IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections [Available online at http://www.ipcc.ch/publications_and_data/publications_and_data_supporting_material.shtml#.Uj7jyd-iJyA, last access: June 10th, 2015].

Koster, R. D., Suarez, M. J., Liu, P., Jambor, U., Berg, A., Kistler, M., Reichle, R., Rodell, M., and Famiglietti, J. Realistic initialization of land surface states: Impacts on subseasonal forecast skill. *Journal of Hydrometeorology*, 5(6):1049–1063, December 2004. ISSN 1525-755X. doi: 10.1175/JHM-387.1.

Koster, R. D., Guo, Z., Yang, R., Dirmeyer, P. A., Mitchell, K., and Puma, M. J. On the Nature of Soil Moisture in Land Surface Models. *Journal of Climate*, 22(16):4322–4335, 2009. ISSN 0894-8755. doi: 10.1175/2009JCLI2832.1.

Koutsoyiannis, D., Kozonis, D., and Manetas, A. A mathematical framework for studying rainfall intensity-duration-frequency relationships. *Journal of Hydrology*, 206(1-2):118–135, 1998.

Kumar, R., Livneh, B., and Samaniego, L. Toward computationally efficient large-scale hydrologic predictions with a multiscale regionalization scheme. *Water Resources Research*, 49(9):5700–5714, 2013b. ISSN 1944-7973. doi: 10.1002/wrcr.20431.

Kumar, R., Samaniego, L., and Attinger, S. The effects of spatial discretization and model parameterization on the prediction of extreme runoff characteristics. *Journal of Hydrology*, 392(1–2):54–69, 2010. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2010.07.047.

Kumar, R., Samaniego, L., and Attinger, S. Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations. *Water Resources Research*, 49(1):360–379, 2013a. ISSN 1944-7973. doi: 10.1029/2012WR012195.

Lafon, T., Dadson, S., Buys, G., and Prudhomme, C. Bias correction of daily precipitation simulated by a regional climate model: a comparison of methods. *International Journal of Climatology*, 2012. ISSN 1097-0088. doi: 10.1002/joc.3518.

Li, H., Sheffield, J., and Wood, E. F. Bias correction of monthly precipitation and temperature fields from Intergovernmental Panel on Climate Change AR4 models using equidistant quantile matching. *Journal of Geophysical Research: Atmospheres*, 115(D10), 2010. ISSN 2156-2202. doi: 10.1029/2009JD012882.

Liang, X., Wood, E. F., and Lettenmaier, D. P. Surface soil moisture parameterization of the vic-2l model: Evaluation and modification. *Global and Planetary Change*, 13(1-4):195 – 206, 1996. ISSN 0921-8181. doi: http://dx.doi.org/10.1016/0921-8181(95)00046-1.

Lopez, A., Fung, F., New, M., Watts, G., Weston, A., and Wilby, R. L. From climate model ensembles to climate change impacts and adaptation: A case study of water resource management in the southwest of England. *Water Resources Research*, 45(8):W08419, 2009. doi: 10.1029/2008WR007499.

Lovejoy, S. and Schertzer, D. Towards a new synthesis for atmospheric dynamics: Space–time cascades. *Atmospheric Research*, 96(1):1–52, 2010. ISSN 0169-8095. doi: 10.1016/j.atmosres.2010.01.004.

Luo, L. and Wood, E. F. Monitoring and predicting the 2007 U.S. drought. *Geophysical Research Letters*, 34(22), 2007. ISSN 1944-8007. doi: 10.1029/2007GL031673.

Luo, L., Wood, E. F., and Pan, M. Bayesian merging of multiple climate model forecasts for seasonal hydrological predictions. *Journal of Geophysical Research: Atmospheres*, 112(D10):n/a–n/a, 2007. ISSN 2156-2202. doi: 10.1029/2006JD007655.

MANOB. MANOB - Mindestanforderung an automatische Niederschlagsmessgeräte (Ombrometer) und deren Betrieb, 2006. Availabe at: http://www.dwd.de/bvbw/generator/DWDWWW/Content/Oeffentlichkeit/KU/KU4/HM/Unsere__Leistungen/Entwicklungsprojekte/manob__pdf,templateId=raw, property=publicationFile.pdf/manob_pdf.pdf; Last access: 7th April 2014.

Maraun, D., Wetterhall, F., Ireson, A. M., Chandler, R. E., Kendon, E. J., Widmann, M., Brienen, S., Rust, H. W., Sauter, T., Themeßl, M., Venema, V. K. C., Chun, K. P., Goodess, C. M., Jones, R. G., Onof, C., Vrac, M., and Thiele-Eich, I. Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Rev. Geophys.*, 48(RG3003):8755–1209, Sep 2010.

Maraun, D. Bias Correction, Quantile Mapping, and Downscaling: Revisiting the Inflation Issue. *Journal of Climate*, 26(6):2137–2143, Jan 2013. ISSN 0894-8755. doi: 10.1175/JCLI-D-12-00821.1.

Matrosov, E., Padula, S., and Harou, J. Selecting Portfolios of Water Supply and Demand Management Strategies Under Uncertainty—Contrasting Economic Optimisation and 'Robust Decision Making' Approaches. *Water Resources Management*, 27(4):1123–1148, 2013. ISSN 0920-4741. doi: 10.1007/s11269-012-0118-x.

Mo, K. C. Drought onset and recovery over the United States. *Journal of Geophysical Research: Atmospheres*, 116(D20), 2011. ISSN 2156-2202. doi: 10.1029/2011JD016168.

Mo, K. C. and Lettenmaier, D. P. Hydrologic prediction over Conterminous U.S. using the National Multi Model ensemble. *Journal of Hydrometeorology*, 2014. ISSN 1525-755X. doi: 10.1175/JHM-D-13-0197.1.

Mo, K. C. and Lyon, B. Global meteorological drought prediction using the north american multi-model ensemble. *Journal of Hydrometeorology*, March 2015. ISSN 1525-755X. doi: 10.1175/JHM-D-14-0192.1.

Mo, K. C., Shukla, S., Lettenmaier, D. P., and Chen, L.-C. Do Climate Forecast System (CFSv2) forecasts improve seasonal soil moisture prediction? *Geophysical Research Letters*, 39(23), 2012. ISSN 1944-8007. doi: 10.1029/2012GL053598.

Music, B. and Caya, D. Evaluation of the Hydrological Cycle over the Mississippi River Basin as Simulated by the Canadian Regional Climate Model (CRCM). *Journal of Hydrometeorology*, 8 (5):969–988, Oct 2007. ISSN 1525-755X. doi: 10.1175/JHM627.1.

Nelsen, R. B. An introduction to copulas. Springer-Verlag New York, 2006. doi: 10.1007/0-387-28678-0.

Olsson, J. Evaluation of a scaling cascade model for temporal rainfall disaggregation . *Hydrology & Earth System Sciences*, 2(1):19–30, 1998.

Pal, J. S., Giorgi, F., Bi, X., Elguindi, N., Solmon, F., a. Rauscher, S., Gao, X., Francisco, R., Zakey, A., Winter, J., Ashfaq, M., Syed, F. S., Sloan, L. C., Bell, J. L., Diffenbaugh, N. S., Karmacharya, J., Konaré, A., Martinez, D., da Rocha, R. P., and Steiner, A. L. Regional Climate Modeling for the Developing World: The ICTP RegCM3 and RegCNET. *Bulletin of the American Meteorological Society*, 88(9):1395–1409, 2007. doi: 10.1175/BAMS-88-9-1395.

Palmer, T. N., Doblas-Reyes, F. J., Hagedorn, R., and Weisheimer, A. Probabilistic prediction of climate using multi-model ensembles: from basics to applications. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360:1991–1998, 2005. doi: 10.1098/rstb.2005.1750.

Paschalis, A., Molnar, P., and Burlando, P. Temporal dependence structure in weights in a multiplicative cascade model for precipitation. *Water Resources Research*, 48(1):n/a–n/a, 2012. ISSN 1944-7973. doi: 10.1029/2011WR010679.

Paschalis, A., Molnar, P., Fatichi, S., and Burlando, P. A stochastic model for high-resolution space-time precipitation simulation. *Water Resources Research*, 2013. ISSN 1944-7973. doi: 10.1002/2013WR014437.

Pelt, S. and Swart, R. Climate Change Risk Management in Transnational River Basins: The Rhine. *Water Resources Management*, 25(14):3837–3861, 2011. ISSN 0920-4741. doi: 10.1007/s11269-011-9891-1.

Piani, C., Haerter, J., and Coppola, E. Statistical bias correction for daily precipitation in regional climate models over Europe. *Theoretical and Applied Climatology*, 99(1-2):187–192, 2010. ISSN 0177-798X. doi: 10.1007/s00704-009-0134-9.

Piayda, A., Dubbert, M., Rebmann, C., Kolle, O., Costa e Silva, F., Correia, A., Pereira, J. S., Werner, C., and Cuntz, M. Drought impact on carbon and water cycling in a mediterranean quercus suber l. woodland during the extreme drought event in 2012. *Biogeosciences*, 11(24):7159–7178, 2014. doi: 10.5194/bg-11-7159-2014.

Radu, R., Déqué, M., and Somot, S. Spectral nudging in a spectral regional climate model. *Tellus A*, 60(5):898–910, 2008.

Rakovec, O., Kumar, R., Mai, J., Schäfer, D., Attinger, S., Cuntz, M., Schrön, M., Thober, S., Zink, M., and Samaniego, L. Multiscale and multivariate evaluation of water fluxes and states over european river basins. submitted to *Journal of Hydrometeorology*, 2015.

Rauscher, S. A., Coppola, E., Piani, C., and Giorgi, F. Resolution effects on regional climate model simulations of seasonal precipitation over Europe. *Climate Dynamics*, 35:685–711, 2010. ISSN 0930-7575. doi: 10.1007/s00382-009-0607-7.

Relief. Horn of africa crisis: 2011-2012, 2011. URL `http://reliefweb.int/disaster/dr-2011-000029-ken`. last access: March 26th, 2015.

Samaniego, L. and Bárdossy, A. Relating macroclimatic circulation patterns with characteristics of floods and droughts at the mesoscale. *Journal of Hydrology*, 335(1-2):109–123, 2007.

Samaniego, L., Kumar, R., and Attinger, S. Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resources Research*, 46(5), 2010. ISSN 1944-7973. doi: 10.1029/2008WR007327.

Samaniego, L., Kumar, R., and Zink, M. Implications of Parameter Uncertainty on Soil Moisture Drought Analysis in Germany. *Journal of Hydrometeorology*, 14(1):47–68, Aug 2013. ISSN 1525-755X. doi: 10.1175/JHM-D-12-075.1.

Samaniego, L., Rakovec, O., Kumar, R., Schaefer, D., Cuntz, M., Mai, J., Thober, S., and Attinger, S. Multiscale prediction and verification of water fluxes and states over large river basins. In *Scientific Program of the AGU General Assembly 2014*, 2014.

Samuelsson, P., Jones, C. G., Willén, U., Ullerstig, A., Gollvik, S., Hansson, U., Jansson, C., Kjellström, E., Nikulin, G., and Wyser, K. The Rossby Centre Regional Climate model RCA3: model description and performance. *Tellus A*, 63(1):4–23, 2011. ISSN 1600-0870. doi: 10.1111/j.1600-0870.2010.00478.x.

Sanchez-Gomez, E., Somot, S., and Déqué, M. Ability of an ensemble of regional climate models to reproduce weather regimes over Europe-Atlantic during the period 1961–2000. *Climate Dynamics*, 33(5):723–736, 2008.

Sansom, P. G., Stephenson, D. B., Ferro, C. A. T., Zappa, G., and Shaffrey, L. Simple Uncertainty Frameworks for Selecting Weighting Schemes and Interpreting Multimodel Ensemble Climate Change Experiments. *Journal of Climate*, 26(12):4017–4037, Jan 2013. ISSN 0894-8755. doi: 10.1175/JCLI-D-12-00462.1.

Schaake, J. C., Hamill, T. M., Buizza, R., and Clark, M. The hydrological ensemble prediction experiment. *Bulletin of the American Meteorological Society*, 88(10):1541–1547, 2007a.

Schaake, J. C., Hamill, T. M., Buizza, R., and Clark, M. Hepex: The hydrological ensemble prediction experiment. *Bulletin of the American Meteorological Society*, 88(10):1541–1547, October 2007b. ISSN 0003-0007. doi: 10.1175/BAMS-88-10-1541.

Schindler, A., Maraun, D., and Luterbacher, J. Validation of the present day annual cycle in heavy precipitation over the British Islands simulated by 14 RCMs. *Journal of Geophysical Research*, 117 (D18107), 2012.

Schuchardt, B., Wittig, S., Mahrenholz, P., Kartschall, K., Mäder, C., Haße, C., and Daschkeit, A. Germany in the midst of climate change: Adaptation is necessary. Federal Environmental Agency, Division I 2.1 ”Climate Change”, P.O. Box 1406, 06813 Dessau-Roßlau, April 2008.

Schulzweida, U. Climate Data Operators User’s Guide, June 2013. [cdo v1.6.1, Available online at https://code.zmaw.de/projects/cdo].

Shahrbanou Madadgar and Hamid Moradkhani. A Bayesian Framework for Probabilistic Seasonal Drought Forecasting. *Journal of Hydrometeorology*, 14(6):1685–1705, 2013. ISSN 1525-755X. doi: 10.1175/JHM-D-13-010.1.

Sheffield, J. and Wood, E. F. Drought: Past Problems and Future Scenarios. Earthscan, 2011. ISBN 9781849710824.

Sheffield, J., Goteti, G., Wen, F., and Wood, E. F. A simulated soil moisture based drought analysis for the United States. *Journal of Geophysical Research: Atmospheres*, 109(D24), 2004. ISSN 2156-2202. doi: 10.1029/2004JD005182.

Sheffield, J., Livneh, B., and Wood, E. F. Representation of terrestrial hydrology and large-scale drought of the continental united states from the north american regional reanalysis. *Journal of Hydrometeorology*, 13(3):856–876, January 2012. ISSN 1525-755X. doi: 10.1175/JHM-D-11-065.1.

Shukla, S. and Lettenmaier, D. P. Seasonal hydrologic prediction in the United States: understanding the role of initial hydrologic conditions and seasonal climate forecast skill. *Hydrology and Earth System Sciences*, 15(11):3529–3538, 2011. doi: 10.5194/hess-15-3529-2011.

Shukla, S., Sheffield, J., Wood, E. F., and Lettenmaier, D. P. On the sources of global land surface hydrologic predictability. *Hydrology and Earth System Sciences*, 17(7):2781–2796, 2013. doi: 10.5194/hess-17-2781-2013.

Shukla, S., McNally, A., Husak, G., and Funk, C. A seasonal agricultural drought forecast system for food-insecure regions of East Africa. *Hydrology and Earth System Sciences*, 18(10):3907–3921, 2014. doi: 10.5194/hess-18-3907-2014.

Sieck, L. C., Burges, S. J., and Steiner, M. Challenges in obtaining reliable measurements of point rainfall. *Water Resources Research*, 43(1), 2007. ISSN 1944-7973. doi: 10.1029/2005WR004519.

Sillmann, J. and Roeckner, E. Indices for extreme events in projections of anthropogenic climate change. *Climatic Change*, 86(1-2):83–104, 2008. doi: 10.1007/s10584-007-9308-6.

Sillmann, J., Kharin, V. V., Zhang, X., Zwiers, F. W., and Bronaugh, D. Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate. *Journal of Geophysical Research: Atmospheres*, 118(4):2473–2493, February 2013. doi: 10.1002/jgrd.50188.

Smith, A. B. and Matthews, J. Quantifying uncertainty and variable sensitivity within the us billion-dollar weather and climate disaster cost estimates. *Natural Hazards*, pages 1–23, 2015. ISSN 0921-030X. doi: 10.1007/s11069-015-1678-x.

Smith, A., Freer, J., Bates, P., and Sampson, C. Comparing ensemble projections of flooding against flood estimation by continuous simulation. *Journal of Hydrology*, 511(0):205–219, 2014. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2014.01.045.

Smith, R. L., Tebaldi, C., Nychka, D., and Mearns, L. O. Bayesian modeling of uncertainty in ensembles of climate models. *Journal of the American Statistical Association*, 104(485):97–116, 2009. doi: 10.1198/jasa.2009.0007.

Soares, P. M. M., Cardoso, R. M., Miranda, P. M. A., Viterbo, P., and Belo-Pereira, M. Assessment of the ENSEMBLES regional climate models in the representation of precipitation variability and extremes over Portugal. *Journal of Geophysical Research: Atmospheres*, 117(D7), 2012. ISSN 2156-2202. doi: 10.1029/2011JD016768.

Suh, M.-S., Oh, S.-G., Lee, D.-K., Cha, D.-H., Choi, S.-J., Jin, C.-S., and Hong, S.-Y. Development of New Ensemble Methods Based on the Performance Skills of Regional Climate Models over South Korea. *Journal of Climate*, 25(20):7067–7082, May 2012. ISSN 0894-8755. doi: 10.1175/JCLI-D-11-00457.1.

Taylor K., Stouffer R., and Meehl G. An Overview of CMIP5 and the Experiment Design. *Bulletin of the American Meteorological Society*, 93(4):485–498, 2012. ISSN 0003-0007. doi: 10.1175/BAMS-D-11-00094.1. doi: 10.1175/BAMS-D-11-00094.1.

Tebaldi, C. and Knutti, R. The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 365(1857):2053–2075, 2007. doi: 10.1098/rsta.2007.2076.

Tebaldi, C., Smith, R. L., Nychka, D., and Mearns, L. O. Quantifying Uncertainty in Projections of Regional Climate Change: A Bayesian Approach to the Analysis of Multimodel Ensembles. *Journal of Climate*, 18(10):1524–1540, May 2005. ISSN 0894-8755. doi: 10.1175/JCLI3363.1.

Thober, S. and Samaniego, L. Robust ensemble selection by multivariate evaluation of extreme precipitation and temperature characteristics. *Journal of Geophysical Research: Atmospheres*, 119 (2):594–613, 2014. ISSN 2169-8996. doi: 10.1002/2013JD020505.

Thober, S., Mai, J., Zink, M., and Samaniego, L. Stochastic temporal disaggregation of monthly precipitation for regional gridded data sets. *Water Resources Research*, 50(11):8714–8735, 2014. ISSN 1944-7973. doi: 10.1002/2014WR015930.

Thober, S., Kumar, R., Sheffield, J., Mai, J., Schäfer, D., and Samaniego, L. On the capability of the north american multi-model ensemble for seasonal soil moisture drought prediction over europe. submitted to *Journal of Hydrometeorology*, 2015.

Twedt, T., Schaake, J. J., and Peck, E. National weather service extended streamflow prediction. In *45th Annual Western Snow Conference*, Albuquerque, New Mexico, April 1977 1977. Western Snow Conference.

Uppala, S. M., K̊Allberg, P. W., Simmons, A. J., Andrae, U., Bechtold, V. D. C., Fiorino, M., Gibson, J. K., Haseler, J., Hernandez, A., Kelly, G. A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R. P., Andersson, E., Arpe, K., Balmaseda, M. A., Beljaars, A. C. M., Berg, L. V. D., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B. J., Isaksen, L., Janssen, P. A. E. M., Jenne, R., Mcnally, A. P., Mahfouf, J.-F., Morcrette, J.-J., Rayner, N. A., Saunders, R. W., Simon, P., Sterl, A., Trenberth, K. E., Untch, A., Vasiljevic, D., Viterbo, P., and Woollen, J. The ERA-40 re-analysis. *Quarterly Journal of the Royal Meteorological Society*, 131(612):2961–3012, 2005. ISSN 1477-870X. doi: 10.1256/qj.04.176.

van den Berg, M. J., Vandenberghe, S., De Baets, B., and Verhoest, N. E. C. Copula-based downscaling of spatial rainfall: a proof of concept. *Hydrology and Earth System Sciences*, 15(5):1445–1457, 2011. doi: 10.5194/hess-15-1445-2011.

van der Linden, P. and Mitchell, J. ENSEMBLES: Climate Change and its Impacts: Summary of research and results from the ENSEMBLES project. Report, Met Office Hadley Centre, FitzRoy Road, Exeter EX1 3PB, UK, 2009.

van Loon, A. F., Tijdeman, E., Wanders, N., van Lanen, H. A. J., Teuling, A. J., and Uijlenhoet, R. How climate seasonality modifies drought duration and deficit. *Journal of Geophysical Research: Atmospheres*, 119(8):4640–4656, 2014. ISSN 2169-8996. doi: 10.1002/2013JD020383.

van Meijgaard, E., van Ulft, L., van de Berg, W., Bosveld, F., van den Hurk, B., Lenderink, G., and Siebesma, A. The KNMI regional atmospheric climate model RACMO version 2.1. Technical report, R.Neth. Meteorol. Inst., De Bilt, The Netherlands, 2008.

Vidal, J.-P., Martin, E., Franchistéguy, L., Habets, F., Soubeyroux, J.-M., Blanchard, M., and Baillon, M. Multilevel and multiscale drought reanalysis over France with the Safran-Isba-Modcou hydrometeorological suite. *Hydrology and Earth System Sciences*, 14(3):459–478, 2010. doi: 10.5194/hess-14-459-2010.

von Storch, H. On the Use of "Inflation" in Statistical Downscaling. *Journal of Climate*, 12(12): 3505–3506, 1999. ISSN 0894-8755. doi: 10.1175/1520-0442(1999)012⟨3505:OTUOII⟩2.0.CO;2.

Wang, A., Lettenmaier, D. P., and Sheffield, J. Soil moisture drought in china, 1950-2006. *Journal of Climate*, 24(13):3257–3271, January 2011. ISSN 0894-8755. doi: 10.1175/2011JCLI3733.1.

Weigel, A. P., Knutti, R., Liniger, M. A., and Appenzeller, C. Risks of Model Weighting in Multimodel Climate Projections. *Journal of Climate*, 23(15):4175–4191, Mar 2010. ISSN 0894-8755. doi: 10.1175/2010JCLI3594.1.

Wilks, D. S. and Wilby, R. L. The weather generation game: a review of stochastic weather models. *Progress in Physical Geography*, 23(3):329–357, 1999. doi: 10.1177/030913339902300302.

Wilks, D. S. A gridded multisite weather generator and synchronization to observed weather data. *Water Resources Research*, 45(10):W10419, Oct 2009. ISSN 0043-1397.

Wilks, D. S. Use of stochastic weathergenerators for precipitation downscaling. *Wiley Interdisciplinary Reviews: Climate Change*, 1(6):898–907, 2010. ISSN 1757-7799. doi: 10.1002/wcc.85.

Wilks, D. S. Statistical Methods in the Atmospheric Sciences. Academic Press, Amsterdam, 3rd edition, 2011. ISBN 9780123850225.

Wilks, D. S. Stochastic weather generators for climate-change downscaling, part II: multivariable and spatially coherent multisite downscaling. *Wiley Interdisciplinary Reviews: Climate Change*, 3(3):267–278, 2012. ISSN 1757-7799. doi: 10.1002/wcc.167.

Wilks, D. Multisite generalization of a daily stochastic precipitation generation model. *Journal of Hydrology*, 210(1–4):178–191, 1998. ISSN 0022-1694. doi: 10.1016/S0022-1694(98)00186-3.

Wood, A. W. and Lettenmaier, D. P. An ensemble approach for attribution of hydrologic prediction uncertainty. *Geophysical Research Letters*, 35(14), 2008. ISSN 1944-8007. doi: 10.1029/2008GL034648.

Wood, A. W., Maurer, E. P., Kumar, A., and Lettenmaier, D. P. Long-range experimental hydrologic forecasting for the eastern United States. *Journal of Geophysical Research: Atmospheres*, 107(D20), 2002. ISSN 2156-2202. doi: 10.1029/2001JD000659.

Yuan, X. and Wood, E. F. Downscaling precipitation or bias-correcting streamflow?: Some implications for coupled general circulation model (CGCM)-based ensemble seasonal hydrologic forecast. *Water Resources Research*, 48(12):W12519, Dec 2012. ISSN 0043-1397. doi: 10.1029/2012WR012256.

Yuan, X. and Wood, E. F. Multimodel seasonal forecasting of global drought onset. *Geophysical Research Letters*, 40(18):4900–4905, 2013. ISSN 1944-8007. doi: 10.1002/grl.50949.

Yuan, X., Wood, E. F., Luo, L., and Pan, M. A first look at Climate Forecast System version 2 (CFSv2) for hydrological seasonal prediction. *Geophysical Research Letters*, 38(13), 2011. ISSN 1944-8007. doi: 10.1029/2011GL047792.

Yuan, X., Wood, E. F., Chaney, N. W., Sheffield, J., Kam, J., Liang, M., and Guan, K. Probabilistic Seasonal Forecasting of African Drought by Dynamical Models. *Journal of Hydrometeorology*, 14 (6):1706–1720, 2013a. ISSN 1525-755X. doi: 10.1175/JHM-D-13-054.1.

Yuan, X., Wood, E. F., Roundy, J. K., and Pan, M. CFSv2-Based Seasonal Hydroclimatic Forecasts over the Conterminous United States. *Journal of Climate*, 26(13):4828–4847, 2013b. ISSN 0894-8755. doi: 10.1175/JCLI-D-12-00683.1.

Yuan, X., Roundy, J. K., Wood, E. F., and Sheffield, J. Seasonal forecasting of global hydrologic extremes: system development and evaluation over GEWEX basins. *Bulletin of the American Meteorological Society*, 2015. ISSN 0003-0007. doi: 10.1175/BAMS-D-14-00003.1.

Zadra, A., Caya, D., Cote, J., Dugas, B., Jones, C., Laprise, R., Winger, K., and Caron, L.-P. The next Canadian Regional Climate Model. *La Physique au Canada*, 64(2):75–83, 2008.

Selbstständigkeitserklärung

Ich erkläre, dass ich die vorliegende Arbeit selbstständig und unter Verwendung der angegebenen Hilfsmittel, persönlichen Mitteilungen und Quellen angefertigt habe.

Leipzig, den 26. Februar 2016

Stephan Thober

Stephan Thober

Lebenslauf

Kanstraße 35
04275 Leipzig
☎ *0341/60453764*
✉ *stephan.thober@gmail.com*

Schulische und akademische Ausbildung

seit Jan 2011 **Doktorand**, *Helmholtz-Zentrum für Umweltforschung*, Leipzig, Abt. Hydrosystemmodellierung.
Thema: Evaluation and Disaggregation of Climate Model Outputs for European Drought Prediction

August 2010 **Erreichen des Grades Diplom-Mathematiker**, *Note 1,2*.
Thema: Averaging Principle for Markov Jump Processes and Applications to Enzyme Kinetics

2005-2010 **Studium**, *Ernst-Moritz-Arndt-Universität Greifswald*, Studiengang Mathematik.
Nebenfach: Physik

Juli 2004 **Erreichen der Allgemeinen Hochschulreife**.

1997-2004 **Besuch des Barnimgymnasium**, *Berlin Hohenschönhausen*.
Leistungskurse: Mathematik und Geschichte

Praktika und Auslandsaufenthalte

Dez. 2007 - Dez. 2008 **Mitglied des Fachschaftsrates Mathematik und Biomathematik**, *Ernst-Moritz-Arndt-Universität Greifswald*.

WS 2009-2010 **Studienaufenthalt im Ausland**, *Bath University*, Bath, United Kingdom.

2008 **Praktikum beim Kompentenzzentrum Wasser Berlin**, *Dauer: 6 Wochen*.

Arbeitserfahrung

10-12/2010 **Studentische Hilfskraft**, *Helmholtz-Zentrum für Umweltforschung*, Leipzig, Abt. Hydrosystemmodellierung.

2005-2006 **Freiwilliges Ökologisches Jahr**.
Träger: Stiftung Naturschutz Berlin, Einsatzstelle: Kompentenzzentrum Wasser Berlin

Publikationen

2015 **Stephan Thober, Rohini Kumar, Justin Sheffield, Juliane Mai, David Schäfer und Luis Samaniego**, On the Capability of the North American Multi-Model Ensemble for Seasonal Soil Moisture Drought Prediction over Europe, Journal of Hydrometeorology, akzeptiert mit wenigen Änderungen.

2015 **Oldrich Rakovec, Rohini Kumar, Juliane Mai, David Schäfer, Sabine Attinger, Matthias Cuntz, Martin Schrön, Stephan Thober, Matthias Zink und Luis Samaniego**, Multiscale and multivariate evaluation of water fluxes and states over European river basins.
Journal of Hydrometeorology, akzeptiert mit umfassenden Änderungen

2015 **Matthias Cuntz, Juliane Mai, Matthias Zink, Stephan Thober, Rohini Kumar, David Schäfer, Martin Schrön, John Craven, Oldrich Rakovec, Diana Spieler, Vladyslav Prykhodko, Giovanni Dalmasso, Jude Musuuza, Ben Langenberg, Sabine Attinger, und Luis Samaniego**, Computationally inexpensive identification of non-informative model parameters by sequential screening, Water Resources Research, akzeptiert mit umfassenden Änderungen.

2014 **Stephan Thober und Luis Samaniego**, Robust Ensemble Selection by Multivariate Evaluation of Extreme Precipitation and Temperature Characteristics, Journal of Geophysical Research: Atmospheres, 119(2), 594-613.

2014 **Stephan Thober, Juliane Mai, Matthias Zink und Luis Samaniego**, Stochastic Temporal Disaggregation of Monthly Precipitation for Regional Gridded Data Sets, Water Resources Research, 50(11), 8714-8735.

Kenntnisse und Interessen

Programmierung	Fortran90, Python, Bash, LaTeX, Office-Programme (Word, Excel)
Sprachen	Deutsch (Muttersprache), Englisch (Verhandlungssicher), Französisch (Grundkenntnisse)
Interessen	Go (ostasiatisches Brettspiel), Gärtnern, Konzertgitarre

Leipzig, 16. Juni 2015

Stephan Thober

HELMHOLTZ
ZENTRUM FÜR
UMWELTFORSCHUNG
UFZ

Helmholtz-Zentrum für Umweltforschung GmbH – UFZ | Permoserstraße 15 | 04318 Leipzig

Prüfungsamt Chemisch-Geowissenschaftliche Fakultät
Friedrich-Schiller-Universität Jena
Humboldtstr. 11
07743 Jena

Stephan Thober
Doktorand
Hydrosystemmodellierung
Tel. ++49 (0)341 235 1050
Fax ++49 (0)341 235 1939
stephan.thober@ufz.de

Leipzig, 18. Juni 2015

Erklärung zu den Eigenanteilen an Publikationen und Zweitpublikationsrechten bei einer kumulativen Publikation sowie die Bestätigung des Einverständnisses der Koautoren

Sehr geehrte Damen und Herren,

hiermit übersende ich Ihnen die Erklärung zu den Eigenanteilen an den drei Publikationen in meiner Dissertation “Evaluation and Disaggregation of Climate Model Outputs for European Drought Prediction”.

Publikation: Robust Ensemble Selection by Multivariate Evaluation of Extreme Precipitation and Temperature Characteristics, *Journal of Geophysical Research: Atmospheres, 119(2):594–613, 2014, ISSN 2169-8996, doi:10.1002/2013JD020505*

	Autor 1: Stephan Thober	Autor 2: Luis Samaniego
Konzeption des Forschungsansatzes	×	×
Planung der Untersuchungen	×	×
Datenerhebung	×	
Datenanalyse und -interpretation	×	×
Schreiben des Manuskripts	×	×
Vorschlag Anrechnung Publikationsäquivalente	1,0	0,75

Publikation: Stochastic Temporal Disaggregation of Monthly Precipitation for Regional Gridded Data Sets, *Water Resources Research, 50(11):8714–8735, 2014, ISSN 1944-7973, doi:10.1002/2014WR015930*

	Autor 1: Stephan Thober	Autor 2: Juliane Mai	Autor 3: Matthias Zink	Autor 4: Luis Samaniego
Konzeption des Forschungsansatzes	×			×
Planung der Untersuchungen	×	×		
Datenerhebung			×	
Datenanalyse und -interpretation	×	×		×
Schreiben des Manuskripts	×	×		×
Vorschlag Anrechnung Publikationsäquivalente	1,0			

Helmholtz-Zentrum für Umweltforschung GmbH – UFZ

Sitz der Gesellschaft: Leipzig

Permoserstraße 15
04318 Leipzig
PF 500136
04301 Leipzig
Tel ++49 (0)341 235-0

info@ufz.de
www.ufz.de

Registergericht: Amtsgericht Leipzig
Handelsregister Nr. B 4703

Vorsitzender des Aufsichtsrats:
MinDirig Wilfried Kraus

Wissenschaftlicher Geschäftsführer:
Prof. Dr. Georg Teutsch

Administrative Geschäftsführerin:
Dr. Heike Graßmann

Bankverbindung:
HypoVereinsbank Leipzig
BLZ 860 200 86
Kto.-Nr. 5080 186 136
Swift(BIC)-Code HYVEDEMM495
IBAN-Nr. DE12860200865080186136
UST-Ident-Nr. DE 141 507 065
Steuer-Nr. 232/124/00416

Zertifikat seit 2014
audit berufundfamilie

EMAS
GEPRÜFTES
UMWELTMANAGEMENT
REG.NR. D-159-00047

Publikation: On the Capability of the North American Multi-Model Ensemble for Seasonal Soil Moisture Drought Prediction over Europe, *submitted to Journal of Hydrometeorology, 2015*

	Autor 1: Stephan Thober	Autor 2: Rohini Kumar	Autor 3: Justin Sheffield	Autor 4: Juliane Mai	Autor 5: David Schäfer	Autor 6: Luis Samaniego
Konzeption des Forschungsansatzes	×		×			×
Planung der Untersuchungen	×	×				×
Datenerhebung	×				×	
Datenanalyse und -interpretation	×	×		×		×
Schreiben des Manuskripts	×	×				×
Vorschlag Anrechnung Publikationsäquivalente	1,0					

Wie aus diesen Tabellen hervorgeht, bin ich an sämtlichen Publikationen als Erstautor beteiligt. Mein Eigenanteil entspricht an jeder einzelnen Publikation mehr als 50%, welches jeweils einem Publikationsäquivalent von 1,0 entspricht.

Ich bestätige hiermit, dass alle Koautoren mit den hier gemachten Angaben zu den Eigenanteilen einverstanden sind.

Darüber hinaus bestätige ich, das die notwendigen Genehmigungen der Verlage für die Zweitpublikation vorliegen.

Mit freundlichen Grüßen,

Stephan Thober

HELMHOLTZ
ZENTRUM FÜR
UMWELTFORSCHUNG
UFZ

Helmholtz-Zentrum für Umweltforschung GmbH – UFZ | Permoserstraße 15 | 04318 Leipzig

Prüfungsamt Chemisch-Geowissenschaftliche Fakultät
Friedrich-Schiller-Universität Jena
Humboldtstr. 11
07743 Jena

Prof. Sabine Attinger
Abteilungsleiterin
Hydrosystemmodellierung
Tel. ++49 (0)341 235 1250
Fax ++49 (0)341 235 1939
sabine.attinger@ufz.de

Leipzig, 15.06.2015

Einverständniserklärung der Betreuerin

Sehr geehrte Damen und Herren,

für alle verwendeten Manuskripte in der kumulativen Dissertation mit dem Titel „Evaluation and Disaggregation of Climate Model Outputs for European Drought Prediction“ von Stephan Thober liegen die notwendigen Genehmigungen der Verlage für die Zweitpublikation vor. Die Koautoren der in dieser kumulativen Dissertation verwendeten Manuskripte sind sowohl über die Nutzung, als auch über die angegebenen Eigenanteile informiert und stimmen diesen zu.

Ich bin mit der Abfassung dieser Dissertation als publikationsbasiert, d.h. kumulativ, einverstanden und bestätige die vorstehenden Angaben. Eine entsprechend begründete Befürwortung mit Angabe des wissenschaftlichen Anteils des Doktoranden an den verwendeten Publikationen werde ich parallel an den Rat der Fakultät der Chemisch-Geowissenschaftlichen Fakultät richten.

Mit freundlichen Grüßen

Sabine Attinger

Helmholtz-Zentrum für Umweltforschung GmbH – UFZ

Sitz der Gesellschaft: Leipzig

Permoserstraße 15
04318 Leipzig
PF 500136
04301 Leipzig
Tel ++49 (0)341 235-0

info@ufz.de
www.ufz.de

Registergericht: Amtsgericht Leipzig
Handelsregister Nr. B 4703

Vorsitzender des Aufsichtsrats:
MinDirig Wilfried Kraus

Wissenschaftlicher Geschäftsführer:
Prof. Dr. Georg Teutsch

Administrative Geschäftsführerin:
Dr. Heike Graßmann

Bankverbindung:
HypoVereinsbank Leipzig
BLZ 860 200 86
Kto.-Nr. 5080 186 136
Swift(BIC)-Code HYVEDEMM495
IBAN-Nr. DE12860200865080186136
UST-Ident-Nr. DE 141 507 065
Steuer-Nr. 232/124/00416

Zertifikat seit 2014
audit berufundfamilie

EMAS
GEPRÜFTES
UMWELTMANAGEMENT
REG.NR. D-159-00047