# Statistical Post-Processing of Ensemble Forecasts: Current Developments and Future Directions

**Tilmann Gneiting**

Heidelberg Institute for Theoretical Studies (HITS)
Karlsruhe Institute of Technology (KIT)

**ECMWF, Reading, United Kingdom**
**February 11, 2015**

**Probabilistic Weather Forecasts**

**Statistical Post-Processing of NWP Ensembles:**
**EMOS/NR and BMA**

**Case Study**

**Accounting for Dependencies:**
**Empirical Copula based Techniques**

**Discussion:**
**Current Developments and Future Directions**

Probabilistic Weather Forecasts

Statistical Post-Processing of NWP Ensembles:

EMOS/NR and BMA

Case Study

Accounting for Dependencies:

Empirical Copula based Techniques

Discussion:

Current Developments and Future Directions

## Probabilistic weather forecasts

**weather forecasting** is considered the **ultimate problem** in meteorology (Bjerknes 1904)

in current practice, medium-range weather forecasting is based on **numerical weather prediction (NWP)** models that represent the physics and chemistry of the atmosphere

however, there are major sources of **uncertainty**, including uncertainty about **initial conditions** and **model parameters**

thus, attention has turned to **probabilistic forecasts**, taking the form of probability distributions over future weather states

preferred approach to probabilistic weather prediction is based on carefully designed **ensembles** of NWP model runs

a global medium-range **ensemble prediction system** has been operational at the **ECMWF** since December 1992 (Buizza and Palmer 1995; Molteni et al. 1996)

## What is a good probabilistic forecast?

Gneiting, Balabdaoui and Raftery (2007) contend that the goal of probabilistic forecasting is to *maximize the sharpness of the predictive distributions subject to calibration*

**calibration**

refers to the **statistical compatibility** between the **predictive distributions** and the verifying **observations**

- joint property of the forecasts and the observations
- in a nutshell, the observations ought to be indistinguishable from samples drawn from the predictive distributions
- can be assessed via **rank** or **probability integral transform (PIT)** histograms

**sharpness**

refers to the **spread** of the **predictive distributions**

- property of the probabilistic forecasts only

## Proper scoring rules

proper scoring rules allow for the **joint** assessment of **calibration** and **sharpness**

a **scoring rule** is a function

$$\mathsf{s}(F, y)$$

that assigns a numerical score to each pair $(F, y)$, where $F$ is the **predictive distribution** and $y$ is the realizing **observation**

we consider scores to be **negatively oriented** penalties that fore-casters aim to **minimize**

a **proper scoring rule** s satisfies the expectation inequality

$$\mathbb{E}_G\, \mathsf{s}(G, Y) \leq \mathbb{E}_G\, \mathsf{s}(F, Y) \quad \text{for all} \quad F, G,$$

thereby encouraging **honest** and **careful** assessments (Gneiting and Raftery 2007)

## Continuous ranked probability score

in meteorological practice, the most popular proper score is the **continuous ranked probability score (CRPS)**,

$$\begin{aligned} \mathrm{s}(F, y) &= \int_{-\infty}^{\infty} (F(x) - \mathbb{1}(x \geq y))^2 \, \mathrm{d}x \\ &= \mathbb{E}_F \, |X - y| - \frac{1}{2} \mathbb{E}_F \, |X - X'| \end{aligned}$$

where $X$ and $X'$ are independent random variables with cumulative distribution function $F$ (Matheson and Winkler 1976; Hersbach 2000; Gneiting and Raftery 2007)

- the CRPS is reported in the **same unit as the observations**
- in the case of a single-valued forecast, the CRPS reduces to the **absolute error**
- thus, the CRPS provides a **direct way** of **comparing single-valued** forecasts and **probabilistic** forecasts

Probabilistic Weather Forecasts

Statistical Post-Processing of NWP Ensembles:

EMOS/NR and BMA

Case Study

Accounting for Dependencies:

Empirical Copula based Techniques

Discussion:

Current Developments and Future Directions

## Statistical Post-Processing of NWP Ensembles

despite their undisputed success, **NWP ensembles** are subject to **model biases** and **lack of calibration**

- in typical experience, rank histograms are U-shaped, indicating **underdispersion** for **surface weather** variables

thus, some form of **statistical post-processing** is required to generate **calibrated** and **sharp** predictive distributions

- idea is to **exploit structure** in past forecast-observation pairs to **correct** for **systematic deficiencies** in the model output
- approach depends on the availability of suitable **training sets**, consisting of past forecast-observation pairs
- typically, a **rolling training period** is used to estimate statistical parameters
- training sets can be usefully **augmented** by **reforecast data**
- simple bias correction doesn't suffice — e.g., in the case of precipitation, additive terms don't work

## EMOS/NR and BMA

two general approaches to the **statistical post-processing** of **NWP ensemble output** have emerged, namely

- **ensemble model output statistics (EMOS)** or **nonhomogeneous regression (NR)**, which fits a single, parametric predictive distribution using summary statistics from the ensemble (Gneiting et al. 2005)

$$y \,|\, x_1, \ldots, x_M \sim f(\, y \,|\, x_1, \ldots, x_M)$$

- **Bayesian model averaging (BMA)**, which fits a mixture density as predictive distribution, where each ensemble member is associated with a kernel function (Raftery et al. 2005)

$$y \,|\, x_1, \ldots, x_M \sim \sum_{m=1}^{M} w_m \, g(y \,|\, x_m)$$

in our experience, the two approaches yield similar predictive performance, with **BMA** being more **flexible**, and **EMOS/NR** being more **parsimonious**

## EMOS/NR and BMA for temperature

consider an **ensemble forecast**, $x_1, \ldots, x_M$, for **temperature**, $y$, at a given time and location

**EMOS/NR** employs a **single Gaussian** predictive density, in that

$$y \,|\, x_1, \ldots, x_M \sim \mathcal{N}(a_0 + a_1 x_1 + \cdots + a_M x_M,\, b_0 + b_1 s^2)$$

with location parameters $b_0$ and $b_1, \ldots, b_M$, and spread parameters $c_0$ and $c_1$, where $s^2$ is the ensemble variance

**BMA** employs Gaussian kernels with a linearly bias-corrected mean, i.e., the BMA predictive density is the **Gaussian mixture**

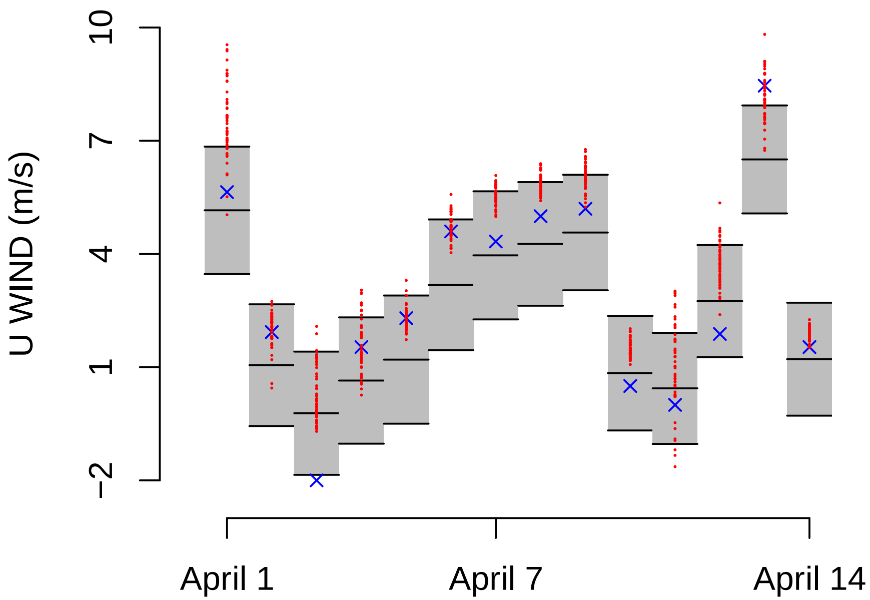
$$y \,|\, x_1, \ldots, x_M \sim \sum_{m=1}^{M} w_m \, \mathcal{N}(c_{0m} + c_{1m} x_m,\, \sigma_m^2)$$

with BMA weights $w_1, \ldots, w_M$, bias parameters $c_{01}, \ldots, c_{0M}$ and $c_{11}, \ldots, c_{1M}$, and spread parameters $\sigma_1^2, \ldots, \sigma_M^2$

for ensembles with groups of **exchangeable** members, such as the ECMWF's ENS, member specific statistical parameters are constrained to be **equal**; e.g., $a_1 = \cdots = a_{50}$ or $w_1 = \cdots = w_{50}$

# Ensemble model output statistics (EMOS) or nonhomogeneous regression (NR)

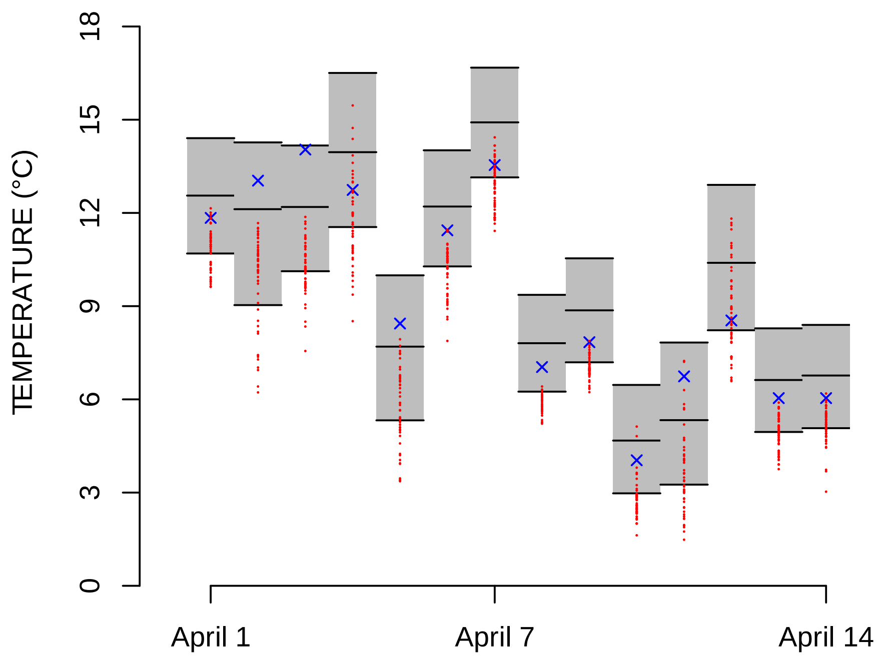
| Weather Quantity | Range | Distribution ( $f$ ) |
|---|---|---|
| Temperature | $y \in \mathbb{R}$ | Normal |
| Pressure | $y \in \mathbb{R}$ | Normal |
| Precipitation amount | $y^{1/2} \in \mathbb{R}^+$ | Truncated logistic |
| | $y \in \mathbb{R}^+$ | Generalized extreme value (GEV) |
| Wind components | $y \in \mathbb{R}$ | Normal |
| Wind speed | $y \in \mathbb{R}^+$ | Truncated normal |
| Cloud cover | $y \in [0, 1]$ | Beta-Bernoulli mixture |




$u$-wind at Hamburg, valid April 1–14, 2011 at 00 UTC, 24-hour lead time

# Bayesian model averaging (BMA)

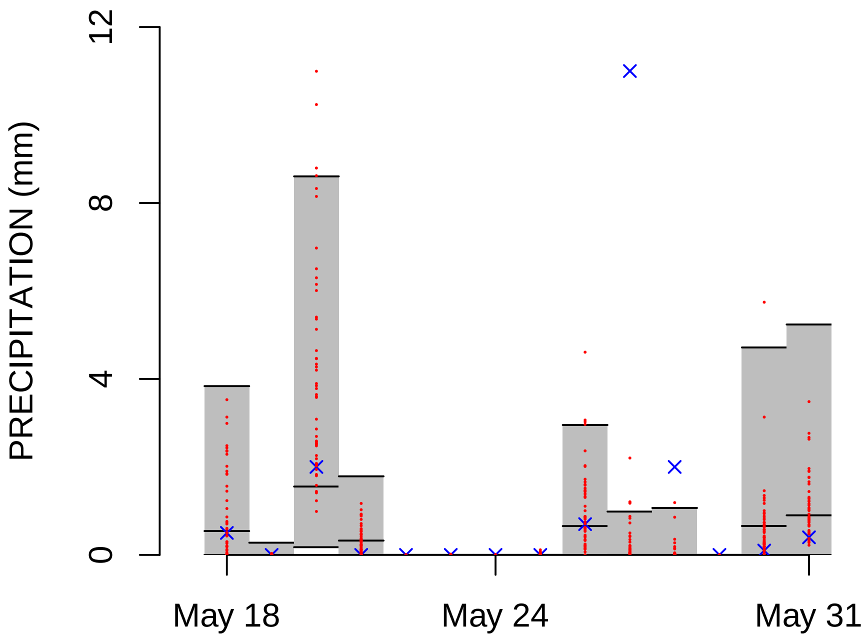
| Variable | Range | Kernel ($g$) | Mean | Variance |
|---|---|---|---|---|
| Temperature | $y \in \mathbb{R}$ | Normal | $c_{0m} + c_{1m} x_m$ | $\sigma_m^2$ |
| Pressure | $y \in \mathbb{R}$ | Normal | $c_{0m} + c_{1m} x_m$ | $\sigma_m^2$ |
| Precipitation accumulation | $y^{1/3} \in \mathbb{R}^+$ | Gamma | $c_{0m} + c_{1m} x_m$ | $d_{0m} + d_{1m} x_m$ |
| Wind components | $y \in \mathbb{R}$ | Normal | $c_{0m} + c_{1m} x_m$ | $\sigma_m^2$ |
| Wind speed | $y \in \mathbb{R}^+$ | Gamma | $c_{0m} + c_{1m} x_m$ | $d_{0m} + d_{1m} x_m$ |
| Visibility | $y \in (0, 1)$ | Beta | $c_{0m} + c_{1m} x_m^{1/2}$ | $d_{0m} + d_{1m} x_m^{1/2}$ |




temperature in Berlin valid April 1–14, 2011 at 00 UTC, 48-hour lead time

# Bayesian model averaging (BMA)

| Variable | Range | Kernel $(f)$ | Mean | Variance |
|---|---|---|---|---|
| Precipitation accumulation | $y^{1/3} \in \mathbb{R}^+$ | Gamma | $c_{0m} + c_{1m} x_m$ | $d_{0m} + d_{1m} x_m$ |




precipitation accumulation in Frankfurt, valid May 18–31, 2011, 24-hour lead time

## EMOS/NR and BMA: selected experience

**EMOS/NR** has been applied to calibrate ensemble forecasts of surface temperature in Austria (Kann et al. 2009, 2011), China, and Germany

Hagedorn (2008) and Hagedorn et al. (2008) report **gains** in lead time **of two to four days** for predictions of surface temperature with the **ECMWF**'s **ENS**, with the improvement being stronger where the original forecast skill is low

Bogner et al. (2013) find gains of about **four days** for surface temperature at Bergen, Vienna, Moscow, Nairobi and Tahiti Island

**BMA** has been applied to calibrate ensemble forecasts of surface temperature in Canada (Wilson et al. 2007), Hungary (Baran et al. 2013), and Iran (Soltanzadeh et al. 2011)

a **real-time implementation** for surface temperature and precipitation over the Pacific Northwest region of the United States is available at `www.probcast.com`

**Probabilistic Weather Forecasts**

**Statistical Post-Processing of NWP Ensembles: EMOS/NR and BMA**

**Case Study**

**Accounting for Dependencies: Empirical Copula based Techniques**

**Discussion: Current Developments and Future Directions**

## Case study

joint work at **ECMWF** and **HITS**

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

**statistical post-processing** for the ECMWF's operational **52-member system** system

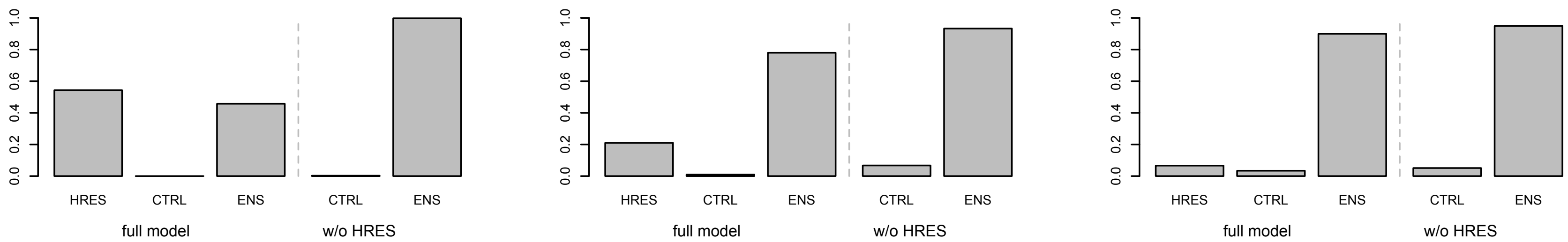
comprising the high-resolution run **(HRES)**, the corresponding 50-member ensemble **(ENS)**, and the control run **(CTRL)**

| Weather Quantity | Acronym | Range | Distribution ($f$) |
|---|---|---|---|
| Temperature | T2M | $y \in \mathbb{R}$ | Normal |
| Precipitation amount | PPT24 | $y \in [0, \infty)$ | Left-censored GEV |
| Wind speed | V10 | $y^{1/2} \in \mathbb{R}^+$ | Truncated normal |
| Cloud cover | TCC | $y \in [0, 1]$ | Beta-Bernoulli mixture |

forecasts with **lead times** from 1 to 10 days, initialized and valid at 12 UTC, respectively

# Training and verification

a **rolling training period** of 1-5 years is used to estimate the EMOS/NR parameters



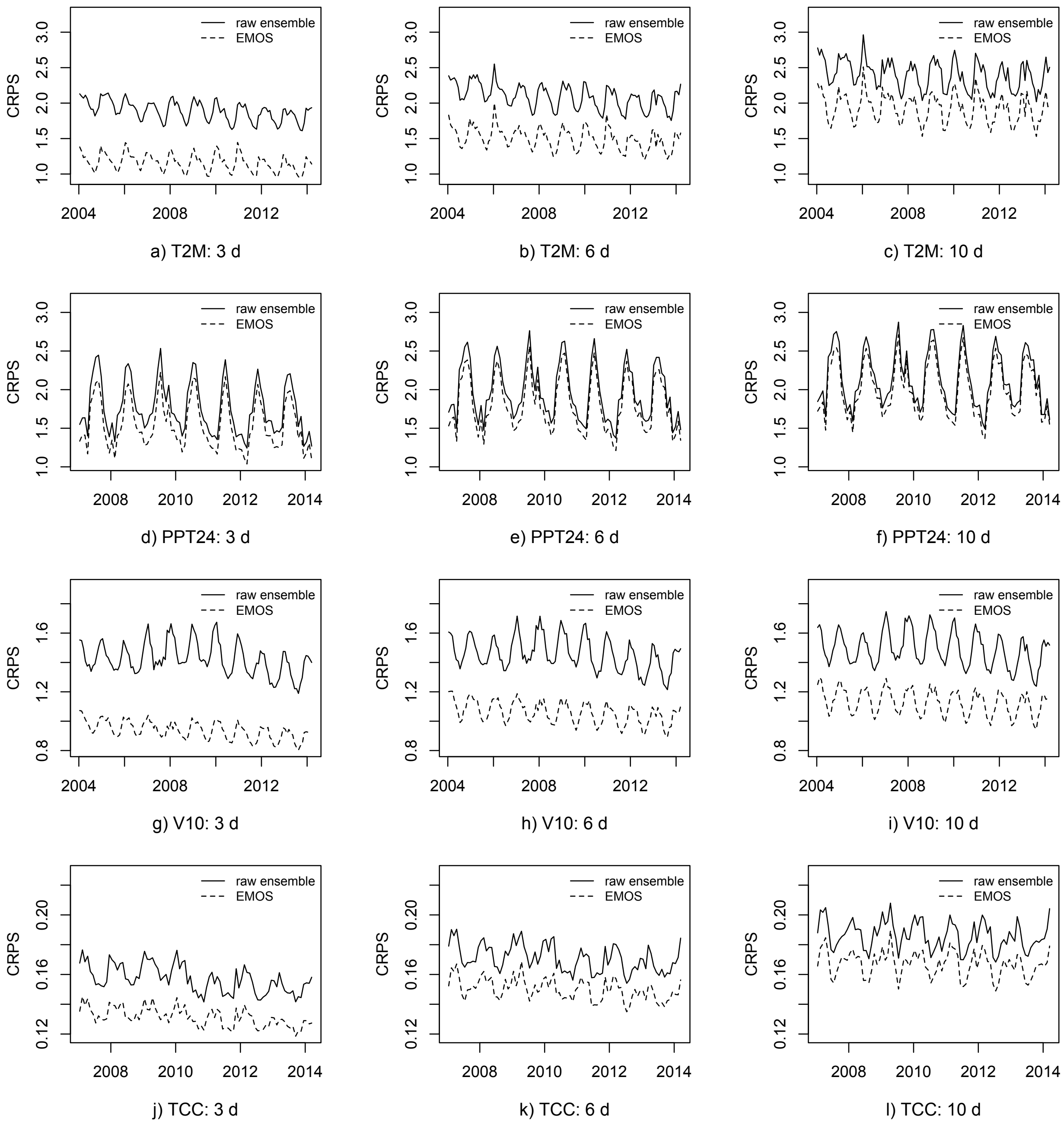

EMOS/NR coefficients for HRES diminish with increasing lead time, as exemplified here for temperature forecasts at Vienna, Austria at lead times of 1, 5, and 10 days

**verification** against thousands of globally distributed **surface synoptic observations (SYNOP)** data

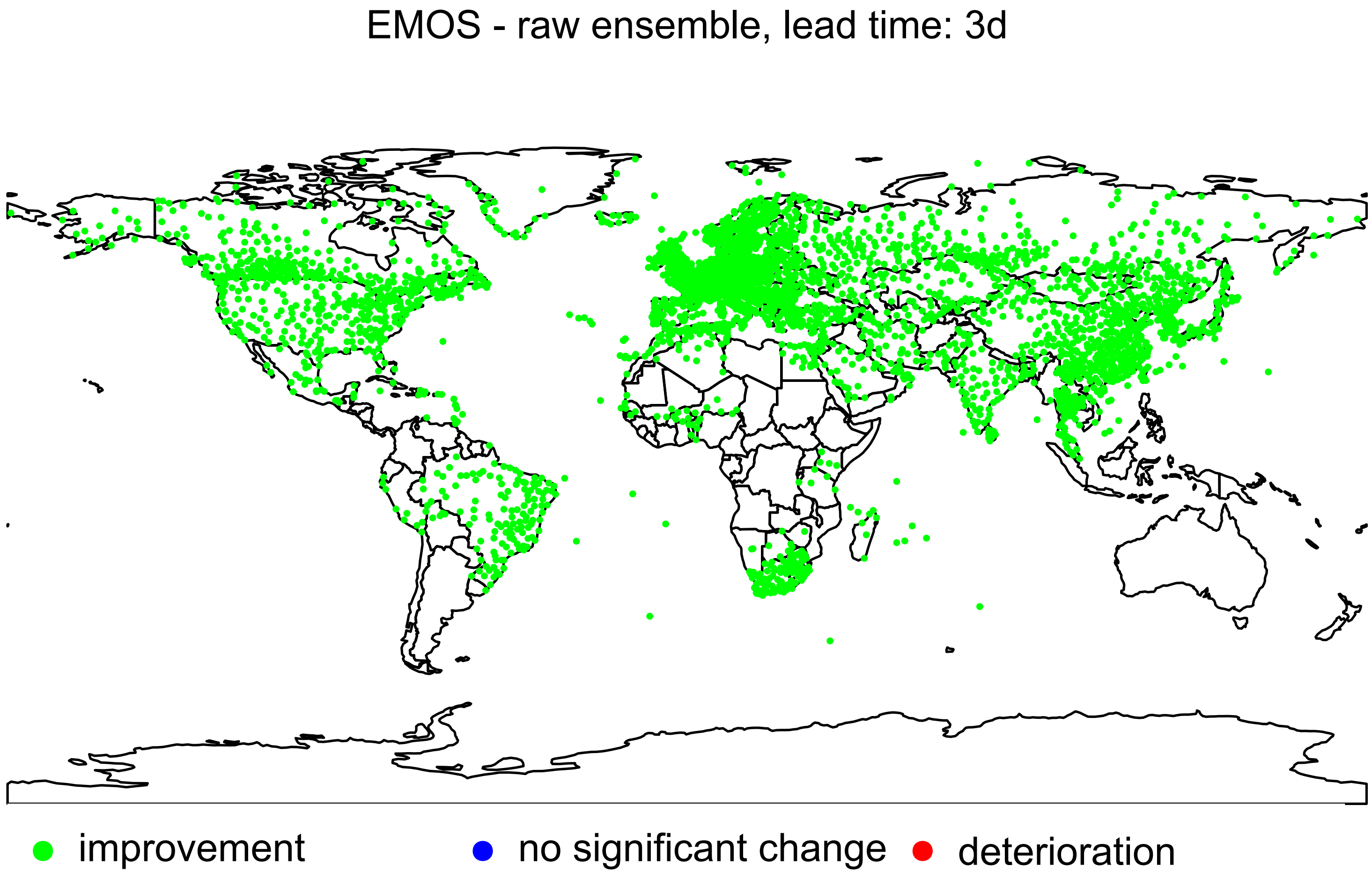
test period from **Jan 1, 2004** to **March 20, 2014**

generally, **statistical post-processing** yields tremendous **improvement** in the predictive performance, as measured by the **CRPS**, especially for T2M, V10, and TCC

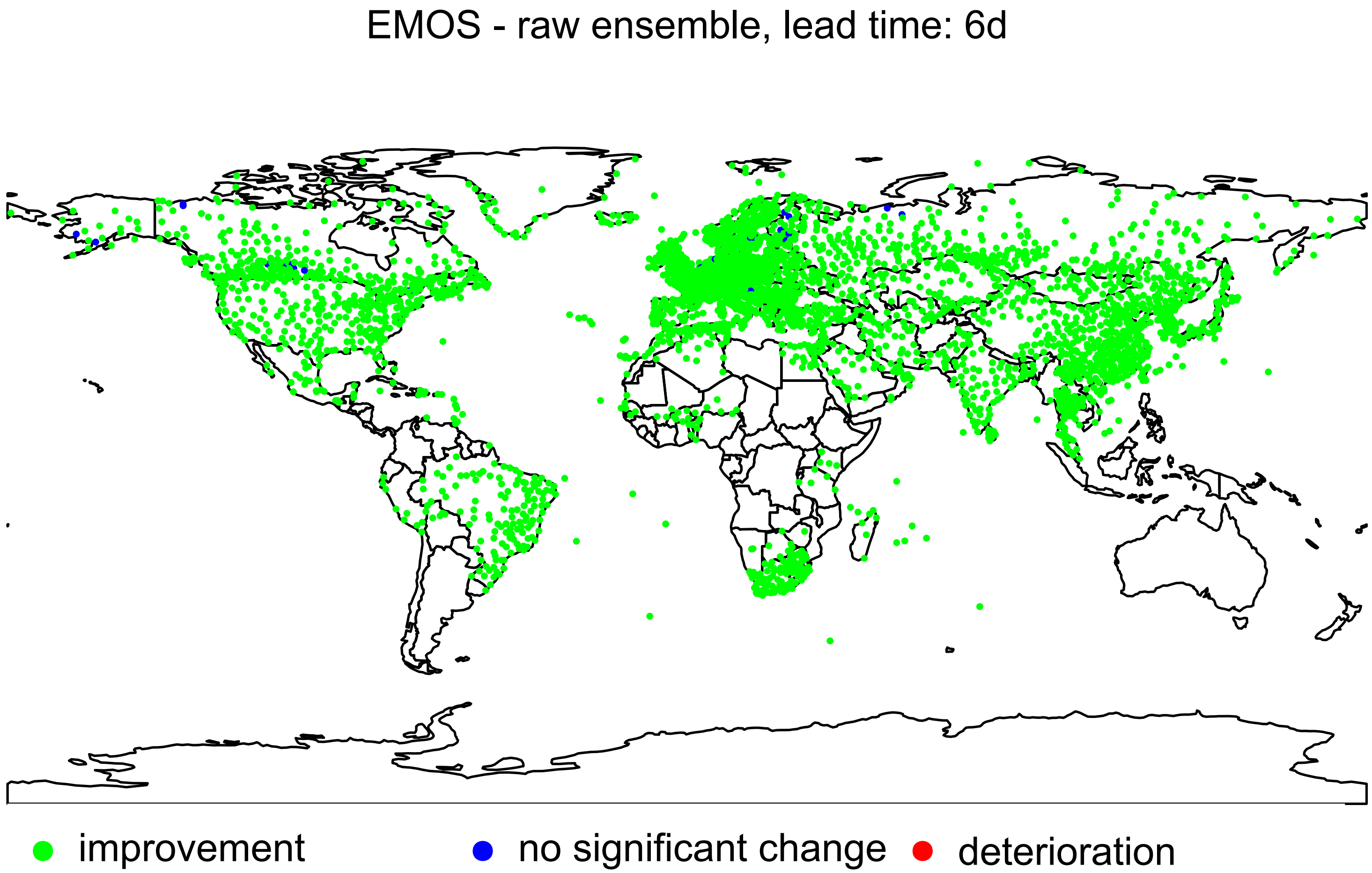
# CRPS: raw ensemble vs. post-processed



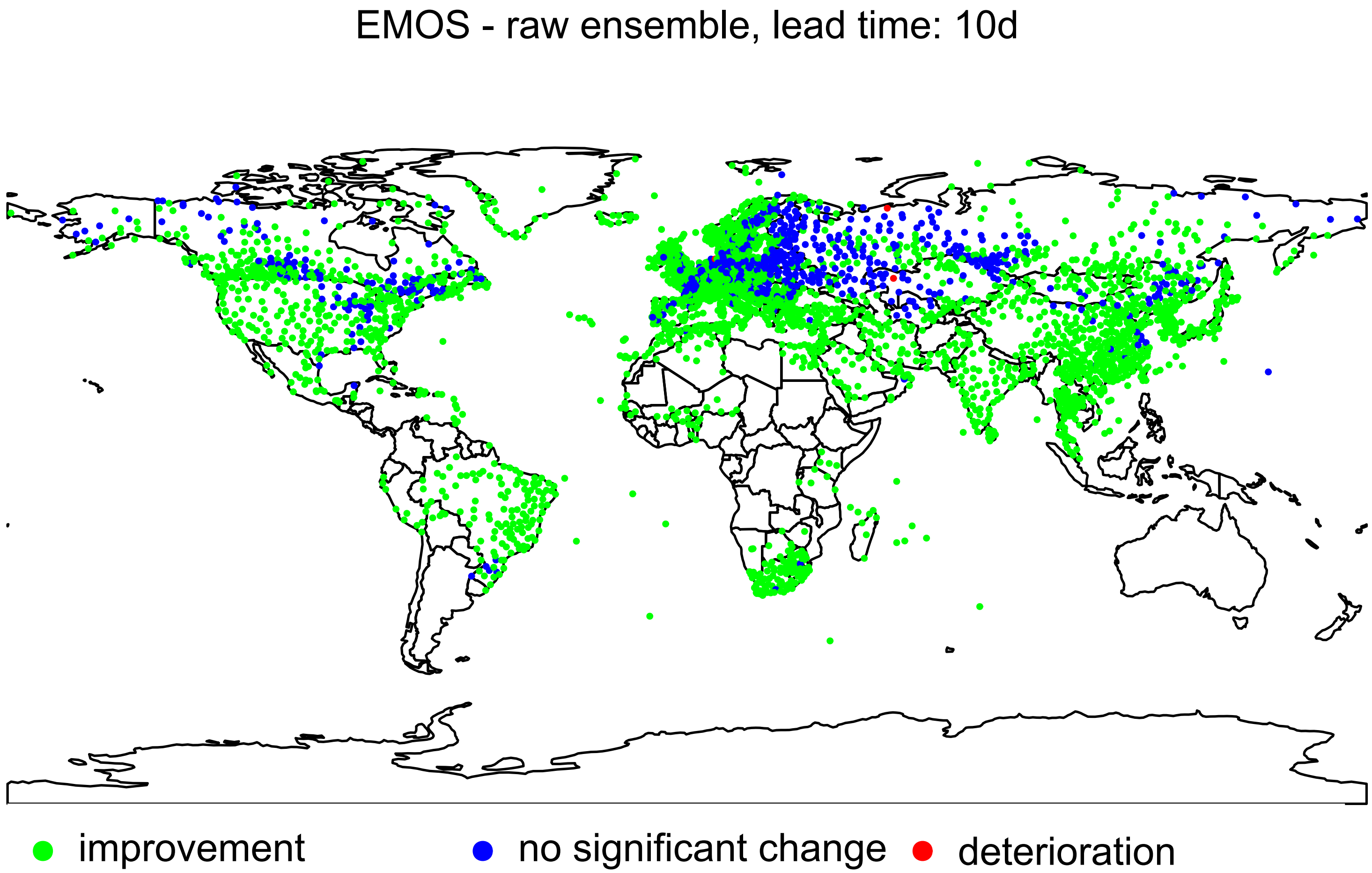
# Temperature: 3 days ahead



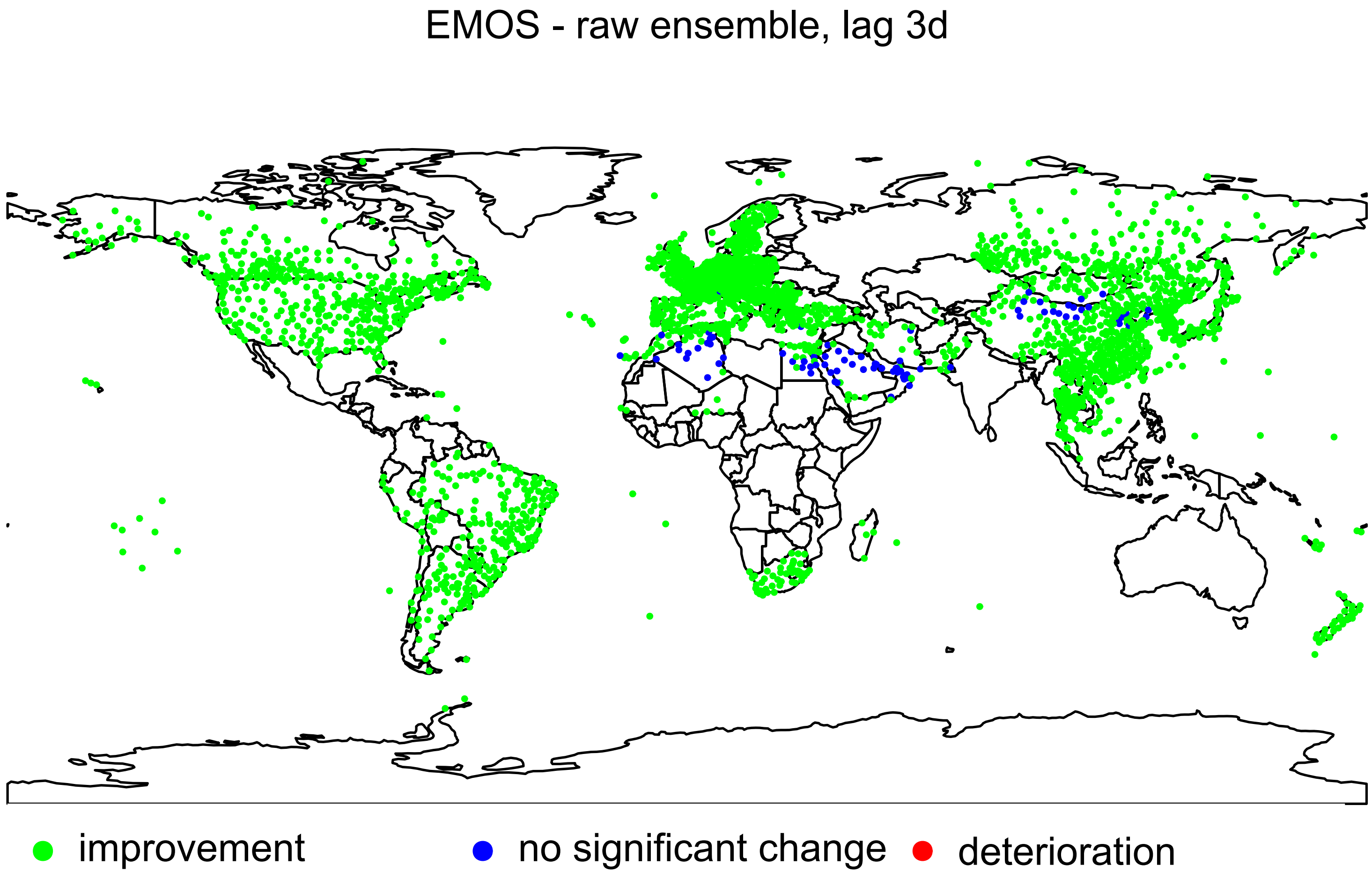
# Temperature: 6 days ahead



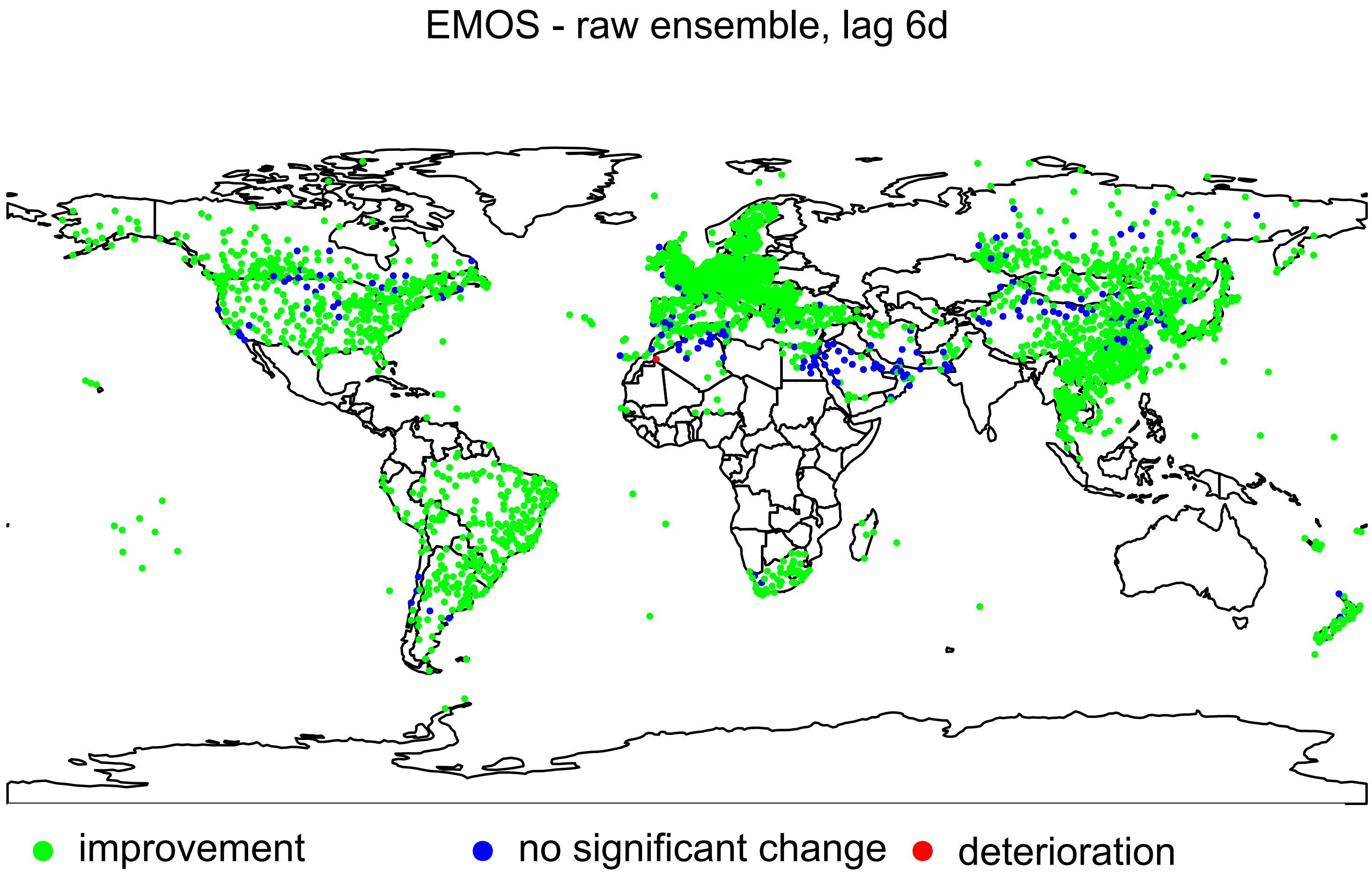
# Temperature: 10 days ahead



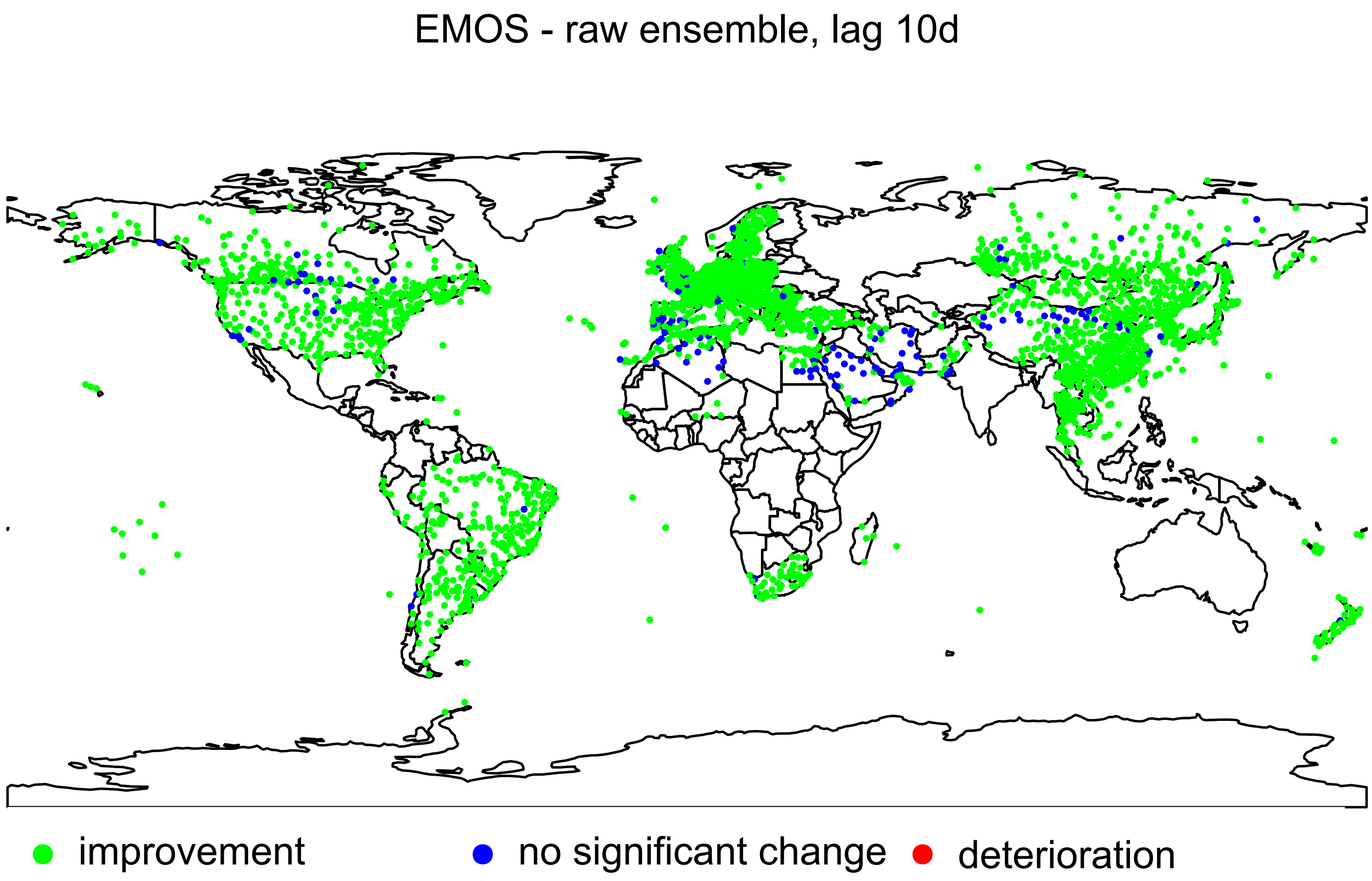
# Precipitation: 3 days ahead



# Precipitation: 6 days ahead



# Precipitation: 10 days ahead



Probabilistic Weather Forecasts

Statistical Post-Processing of NWP Ensembles:
EMOS/NR and BMA

Case Study

Accounting for Dependencies:
Empirical Copula based Techniques

Discussion:
Current Developments and Future Directions

## Accounting for dependencies

EMOS/NR and BMA apply to any **single weather variable** at any **single location** and any **single look-ahead time**

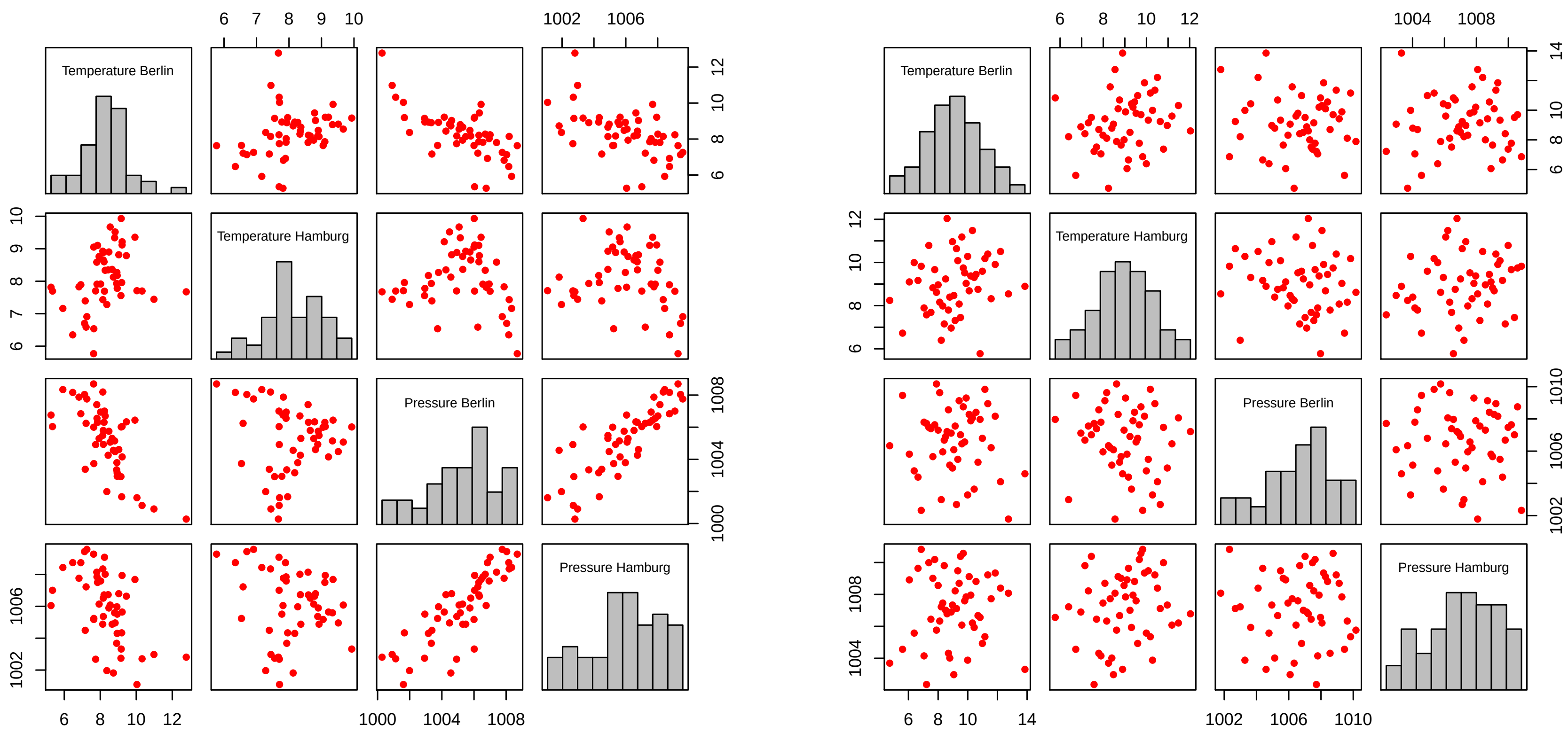
however, individually post-processed distributions fail to account for **multivariate dependence** structures

the most pressing need now is to develop post-processing techniques that yield **physically realistic** probabilistic forecasts of **spatio-temporal weather trajectories** for **multiple weather variables** at **multiple locations** and **multiple look-ahead times**

key applications include **air traffic control**, **ship routeing**, and **hydrologic** predictions

# Example

illustration: 24-hour **ECMWF ENS** forecast of surface temperature and pressure at Berlin and Hamburg valid May 27, 2010 **before** and **after** BMA **post-processing**



## Sklar's Theorem

EMOS/NR and BMA apply to any **single weather variable** at any **single location** and any **single look-ahead time**

yielding a **marginal** predictive **cumulative distribution function (CDF)**, $F_l$, for any given **univariate** weather quantity, $Y_l$

with each multi-index $l = (i, j, k)$ referring to **weather variable** $i$, **location** $j$, and **look-ahead time** $k$

we seek a **physically realistic** and consistent **multivariate** or **joint** predictive **CDF**, $F$, with **margin** $F_l$ for each $l = 1, \ldots, L$

**Sklar's theorem (1959):** every multivariate CDF $F$ with margins $F_1, \ldots, F_L$ can be written as

$$F(y_1, \ldots, y_L) = C(F_1(y_1), \ldots, F_L(y_L))$$

where $C : [0, 1]^L \to [0, 1]$ is a **copula**, i.e., a multivariate CDF with standard **uniform margins**

## Copula approaches

in order to issue **physically realistic** and **consistent** probabilistic forecasts of **spatio-temporal weather trajectories**

it remains to specify and fit a suitable **copula** $C : [0,1]^L \to [0,1]$

if $L$ is small, or if specific structure can be exploited, **parametric** families of copulas work well

- Gel et al. (2004), Berrocal et al. (2007), Pinson et al. (2009), Schuhen et al. (2012) and Möller et al. (2013) use **Gaussian copulas**
- parametric or semi-parametric alternatives include **elliptical**, **Archimedean**, hierarchical Archimedean and **pair** copulas

if $L$ is huge and no specific structure can be exploited, we need to resort to **non-parametric** approaches, based on **empirical copulas**, with the **Schaake shuffle** (Clark et al. 2004) and **ensemble copula coupling (ECC)** being particularly attractive options

## Ensemble copula coupling (ECC; Schefzik et al. 2013)

given an **NWP ensemble** of size $M$ for the weather variables $Y_l$, where $l = 1, \ldots, L$, **ECC** proceeds in three steps

**univariate post-processing:** for each $l = 1, \ldots, L$, apply **EMOS/NR** or **BMA** to obtain a post-processed **predictive CDF**, $F_l$

**quantization:** for each $l = 1, \ldots, L$, obtain a discrete **sample** of size $M$ from $F_l$, e.g., using

$$\tilde{x}_m = F_l^{-1}\left(\frac{m}{M+1}\right), \quad m = 1, \ldots, M$$

**ensemble reordering:** take the function $C : [0,1]^L \to [0,1]$ in Sklar's theorem to be the **empirical copula** of the raw ensemble, i.e., arrange the post-processed values in the same **rank order** as the **raw ensemble** values

# Ensemble copula coupling (ECC)

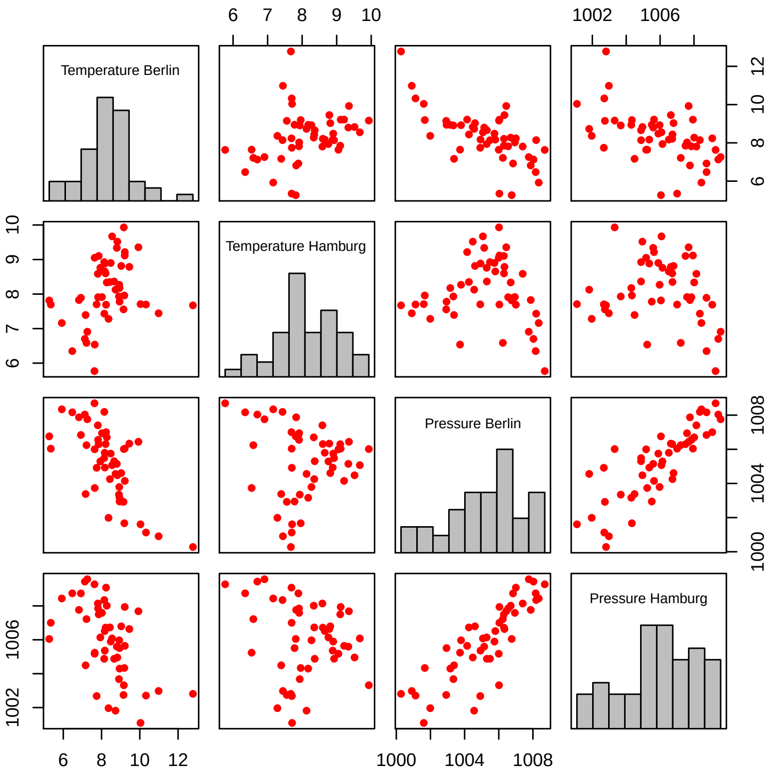
the method is **implicit** or **explicit** in scattered **recent work**, including that of Bremnes (2007), Krzysztofowicz and Toth (2008), Pinson (2011), Flowerdew (2012), Roulin and Vannitsem (2012) and Schuhen, Thorarinsdottir and Gneiting (2012)

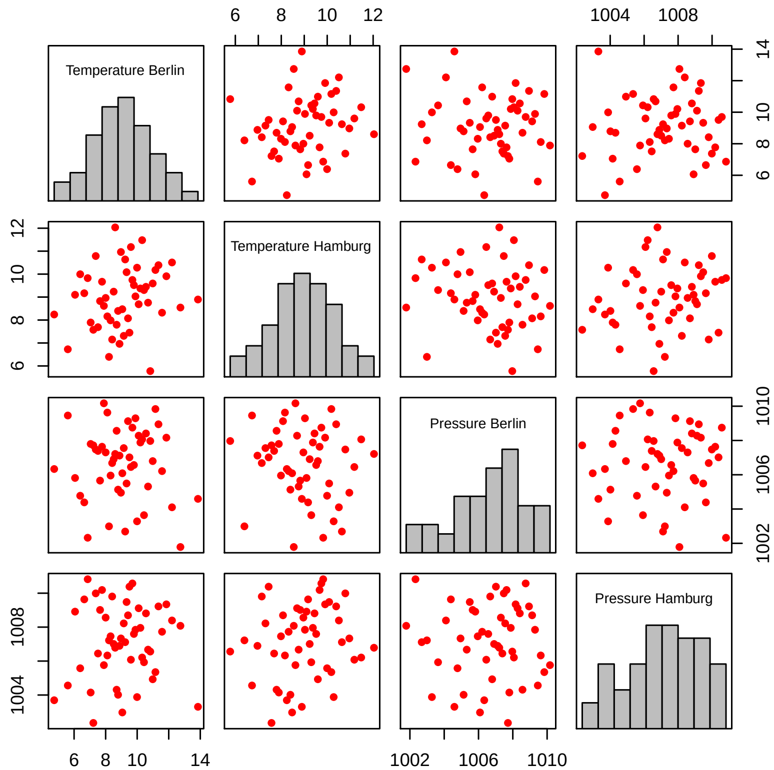
Flowerdew (2012, p. 15) explains colorfully:

> The key to preserving spatial, temporal and inter-variable structure is how this set of values is distributed between ensemble members. One can always construct ensemble members by sampling from the calibrated PDF, but this alone would produce spatially noisy fields lacking the correct correlations. Instead, *the values are assigned to ensemble members in the same order as the values from the raw ensemble: the member with the locally highest rainfall remains locally highest, but with a calibrated rainfall magnitude.*

# Ensemble copula coupling (ECC)

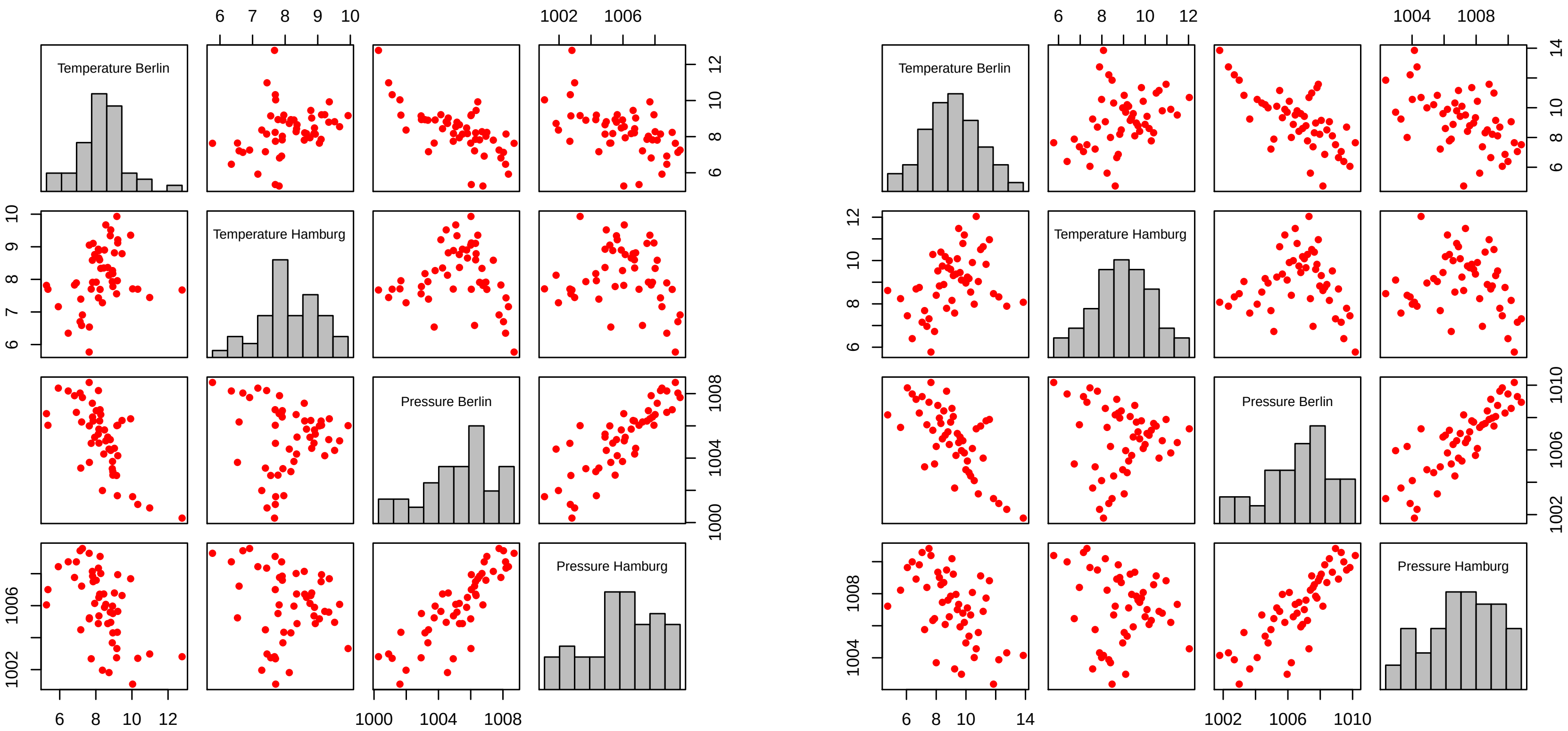
illustration: 24-hour **ECMWF ENS** forecast of surface temperature and pressure at Berlin and Hamburg valid May 27, 2010 before and after post-processing with **BMA**

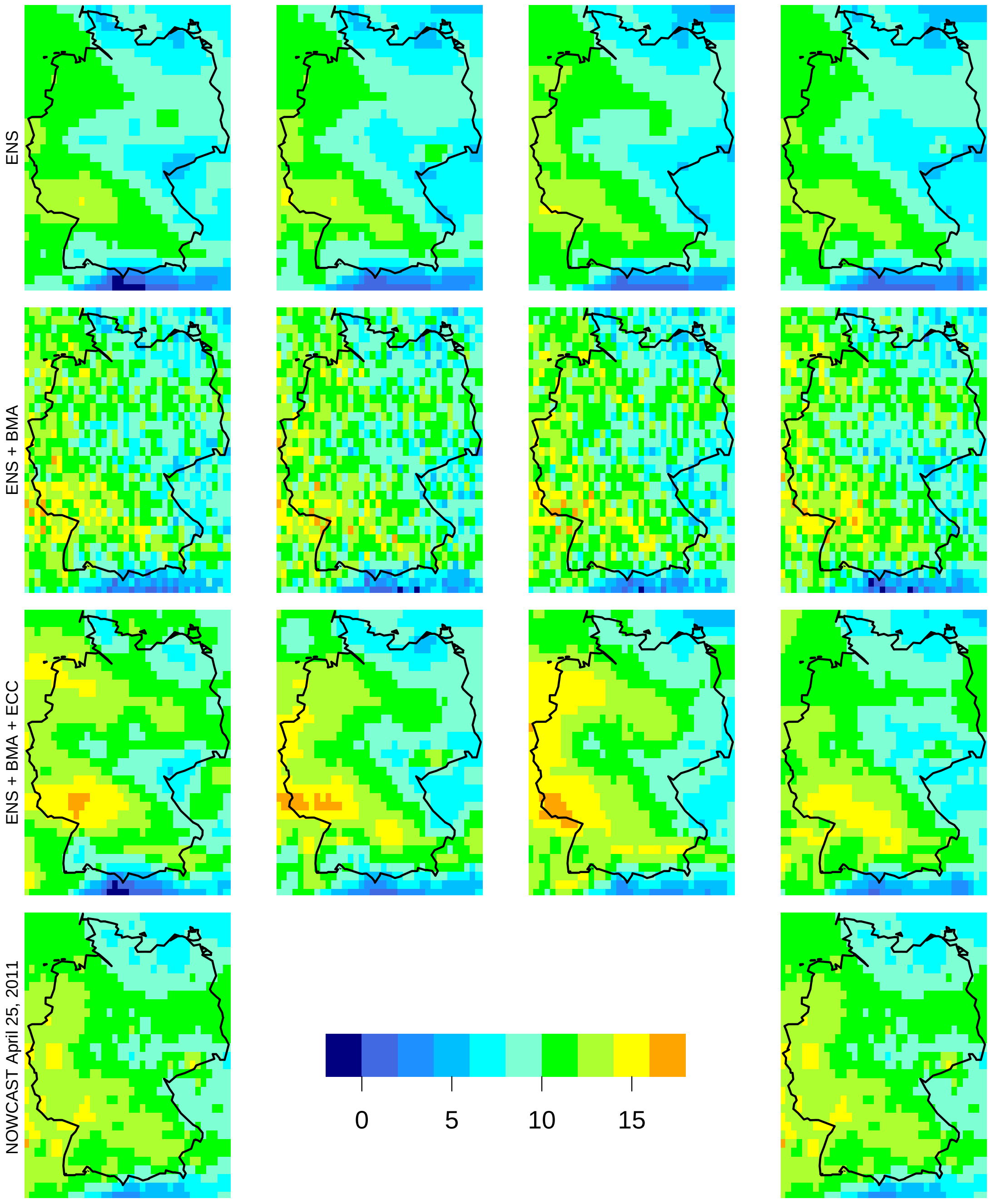


# Ensemble copula coupling (ECC)

illustration: 24-hour **ECMWF ENS** forecast of surface temperature and pressure at Berlin and Hamburg valid May 27, 2010 before and after post-processing with **BMA + ECC**




ENS
ENS + BMA
ENS + BMA + ECC
NOWCAST April 25, 2011
0
5
10
15

## Ensemble copula coupling (ECC; Schefzik et al. 2013)

given an **NWP ensemble** of size $M$ for the weather variables $Y_l$, where $l = 1, \ldots, L$, **ECC** proceeds in three steps

**univariate post-processing:** for each $l = 1, \ldots, L$, apply **EMOS/NR** or **BMA** to obtain a post-processed **predictive CDF**, $F_l$

**quantization:** for each $l = 1, \ldots, L$, obtain a discrete **sample** of size $M$ from $F_l$, e.g., using

$$\tilde{x}_m = F_l^{-1}\left(\frac{m}{M+1}\right), \quad m = 1, \ldots, M$$

**ensemble reordering:** take the function $C : [0,1]^L \to [0,1]$ in Sklar's theorem to be the **empirical copula** of the raw ensemble, i.e., arrange the post-processed values in the same **rank order** as the **raw ensemble** values

## Schaake shuffle (Clark et al. 2004)

given an **NWP ensemble** of size $M$ for the weather variables $Y_l$, where $l = 1, \ldots, L$, the **Schaake shuffle** proceeds in three steps

**univariate post-processing:** for each $l = 1, \ldots, L$, apply **EMOS/NR** or **BMA** to obtain a post-processed **predictive CDF**, $F_l$

**quantization:** for each $l = 1, \ldots, L$, obtain a discrete **sample** of size $N$ from $F_l$, e.g., using

$$\tilde{x}_n = F_l^{-1}\left(\frac{n}{N+1}\right), \quad n = 1, \ldots, N$$

**ensemble reordering:** take the function $C : [0,1]^L \to [0,1]$ in Sklar's theorem to be the **empirical copula** of a relevant historical weather record of size $N$, i.e., arrange the post-processed values in the same **rank order** as the **weather record**

**Probabilistic Weather Forecasts**

**Statistical Post-Processing of NWP Ensembles:**
**EMOS/NR and BMA**

**Case Study**

**Accounting for Dependencies:**
**Empirical Copula based Techniques**

**Discussion:**
**Current Developments and Future Directions**

## Current developments

the basic idea of **statistical post-processing** is to exploit structure in past forecast-observation pairs to **correct** for **probabilistic biases** in NWP model output

for the ECMWF ensemble, **ensemble model output statistics (EMOS/NR)** or **Bayesian model averaging (BMA)** yield **gains** in lead time **of several days** for forecasts of **surface weather**

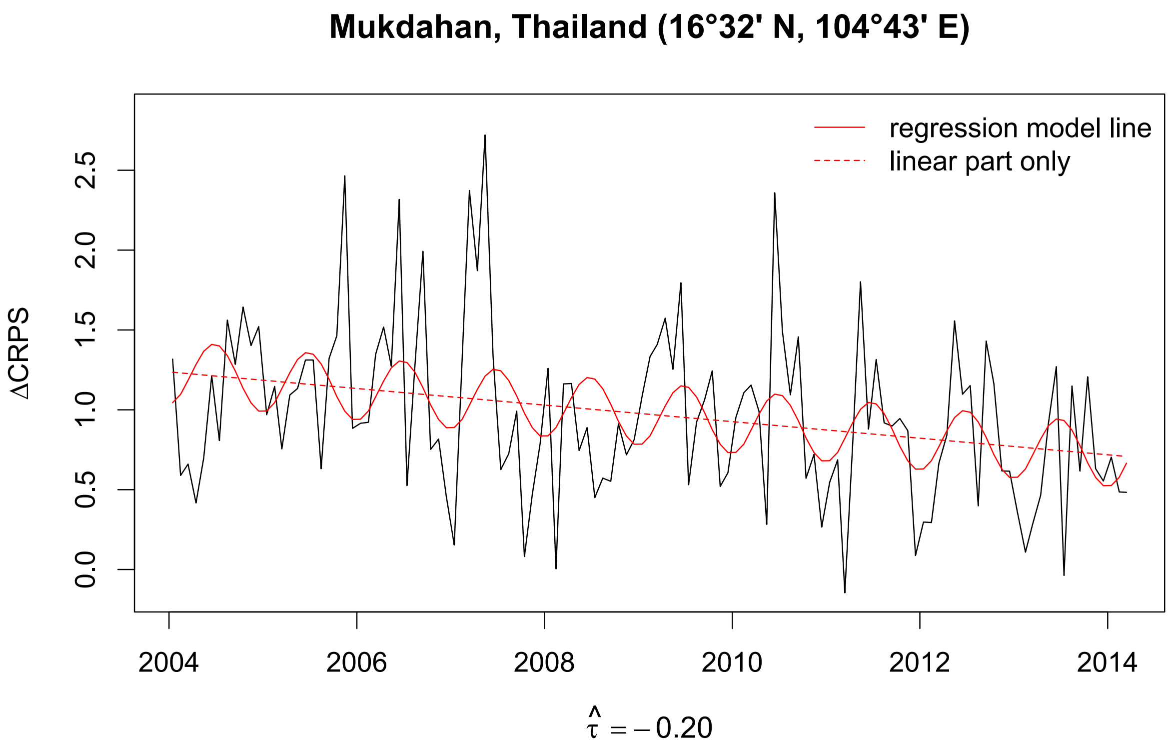
> with the improvement being strongest where the original forecast skill is low, such as in regions with complex terrain

however, as raw ensembles continue to improve, a natural **hypothesis** is that the **gap in skill** between raw ensembles and statistically post-processed forecasts **narrows**

> to investigate this, Hemri et al. (2014) study the evolution of the skill gap between 2004 and 2014

# Trends in predictive performance

Hemri et al. (2014) study the **evolution** of the **skill gap** between the ECMWF raw ensemble and EMOS/NR post-processed forecasts, as measured by monthly **$\Delta$CRPS**



# Trends in predictive performance

percentage of stations with a negative, no significant, or a positive **trend** in monthly **$\Delta$CRPS**

| Lead | Trend | T2M | PPT24 |
|---|---|---|---|
| 3 days | negative | 32% | 18% |
| | not significant | 44% | 77% |
| | positive | 24% | 5% |
| 6 days | negative | 29% | 13% |
| | not significant | 48% | 82% |
| | positive | 23% | 5% |
| 10 days | negative | 26% | 11% |
| | not significant | 54% | 82% |
| | positive | 20% | 7% |

the skill gap tends to remain constant over time, suggesting that post-processing will keep adding value in the foreseeable future

# Future directions

**statistical parameters** need to be **estimated** from **training data**, which can be usefully augmented by using **reforecast data**

> however, it is not obvious how to optimally design an operational reforecast system (Hagedorn 2008; Hamill et al. 2008) and adapt estimation strategies (Roulin and Vannitsem 2012)

handling of **extreme events** in statistical post-processing ought to be studied further

> we rely on the NWP model's ability to signal pending extreme events
>
> generalized extreme value (GEV) distributions have been employed in EMOS/NGR approaches for peak wind and precipitation (Friederichs and Thorarinsdottir 2012; Scheuerer 2014)
>
> a related approach employs GEV distributions only if the ensemble median is above a high threshold (Lerch and Thorarinsdottir 2013)
>
> there is a pronounced need for further methodological development and comparative studies (e.g., Williams et al. 2014)

## Future directions

much current research and development focuses on post-processing techniques for **multiple weather variables** at **multiple locations** and **multiple look-ahead times** simultaneously

with the goal of generating calibrated and sharp ensemble forecasts of spatio-temporal weather scenarios

**empirical copula** based approaches such as **ensemble copula coupling (ECC)** or the **Schaake shuffle** show promise

ECC adopts the rank dependence structure from the ensemble forecast: a purely model based approach

depending on the details of the quantization step, we distinguish ECC-Q, ECC-R, and ECC-T (Schefzik et al. 2013)

the Schaake shuffle adopts the rank dependence structure from a historical weather record: a purely data based approach

it is not obvious how to select historical weather observations that are relevant to the ensemble forecast at hand

there is a pronounced need for further methodological development and comparative studies (e.g., Wilks 2014)

## Selected references


Gneiting, T. (2014). Calibration of Medium-Range Weather Forecasts. ECMWF Technical Memorandum **719**.

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

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

Schefzik, R., Thorarinsdottir, T. L. and Gneiting, T. (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, **28**, 616–640.

Probabilistic Weather Forecasts

Statistical Post-Processing of NWP Ensembles:
EMOS/NR and BMA

Case Study

Accounting for Dependencies:
Empirical Copula based Techniques

Discussion:
Current Developments and Future Directions