D-Vine Copula based Postprocessing of Wind Speed Ensemble Forecasts

David Jobst [1] Annette Möller [2] Jürgen Groß[1]

[1]University of Hildesheim, Institute of Mathematics and Applied Informatics [2]Bielefeld University, Faculty of Business Administration and Economics

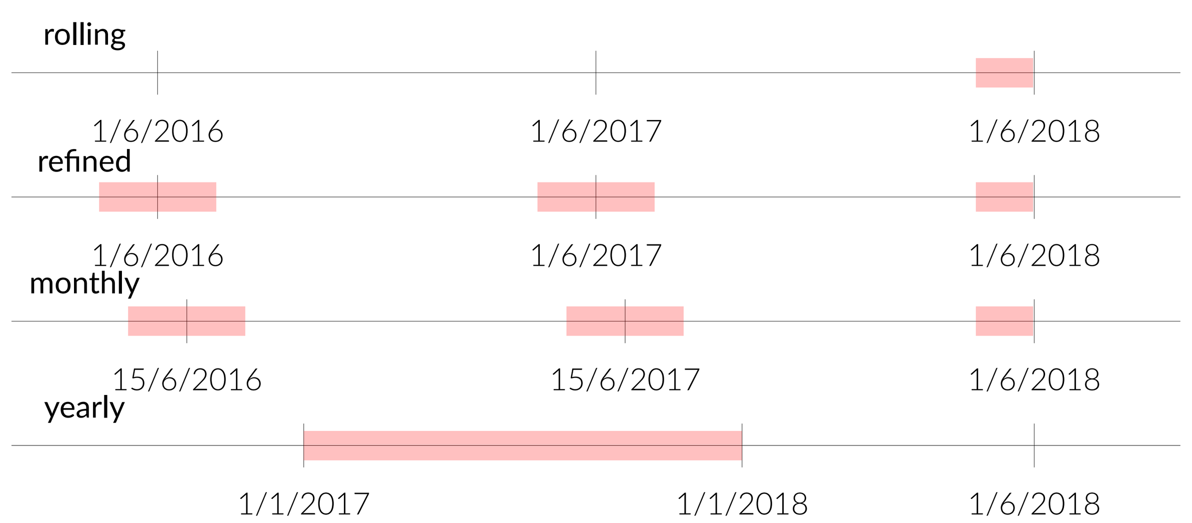
EGU General Assembly #EGU23 23 - 28 April 2023

1. Introduction

- Ensemble forecasts are biased and uncalibrated (underdispersive).
- Ensemble postprocessing (PP) models are used to correct the systematic errors and to calibrate the forecasts.
- These models often implicitly assume a Gaussian dependence between the weather variable y and the ensemble forecasts $x_1, \ldots, x_m$.
- Motivation: Allow a flexible dependence between y and $x_1, \ldots, x_m$ by a vine copula.

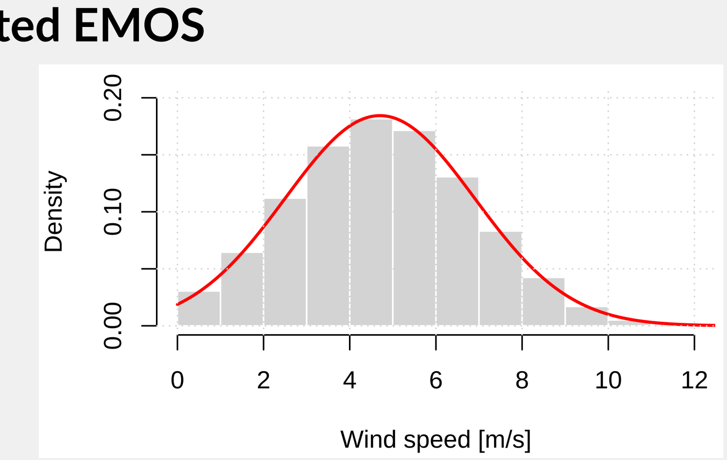
2. Data

- Surface wind speed observations y from 60 SYNOP stations in Germany (German Weather Service).
- ECMWF ensemble mean $\overline{x}_\bullet$ and control $x_\bullet^{\text{ctrl}}$ forecasts with 24 h leadtime of nine weather variables: surface wind speed (wspd), surface wind speed u- and v-component (u,v), surface wind direction (angle), surface wind gust (wgust), surface temperature (temp), surface pressure (sp), total cloud cover (tcc), specific humidity at 850hPa (sh).
- Notation: For example ensemble mean $\overline{x}_{\text{wspd}}$ and control $x_{\text{wspd}}^{\text{ctrl}}$ forecast of surface wind speed (wspd).
- Three spatial variables: station longitude and latitude ($x_{\text{lon}}, x_{\text{lat}}$), station elevation ($x_{\text{elev}}$).
- Training data: January 2, 2016 to May 31, 2018; Validation data: June 1, 2018 to December 31, 2020.
- Four types of training periods: For forecasting, e.g. 1/6/2018 we use the following model training data:



3. Reference Model

Zero-truncated EMOS



- Zero-truncated ensemble model output statistics (tEMOS) as reference model, i.e.

$$y \sim \mathcal{N}_0(\mu, \sigma^2),$$

$$\mu(t) := a_0 + a_1 \overline{x}_{\text{wspd}}(t) + a_2 x_{\text{wspd}}^{\text{ctrl}}(t), \quad \sigma^2(t) := b_0 + b_1 S_{\text{wspd}}^2(t),$$

with ensemble mean $\overline{x}_{\text{wspd}}$, control forecast $x_{\text{wspd}}^{\text{ctrl}}$ and empirical ensemble variance S_{wspd}^2.
- Parameter optimization with respect to CRPS (continuous ranked probability score).

4. D-Vine Copula based PP

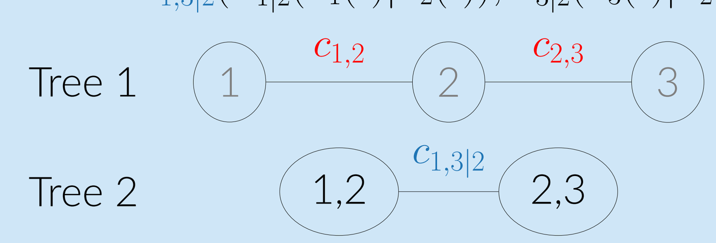
Model description

- A p-dimensional copula is a multivariate distribution function on $[0,1]^p$.
- Sklar's Theorem: For a p-dimensional distribution function, there exists a copula C, such that

$$F(x_1(t), \ldots, x_p(t)) = C(F_1(x_1(t)), \ldots, F_p(x_p(t))),$$

where $F_\bullet$ are the marginals of $x_\bullet$.
- Standard multivariate copulas C are too inflexible for dependence modelling.
- Solution: Vine copulas, where the joint dependence C is build up by only bivariate copulas (pair-copula construction) and a nested set of trees (vine).
- D-vine: Nested set of trees in which each tree is a path.
- Example for a $p = 3$-dimensional D-vine with its pair-copulas densities $c_\bullet$ and the joint density f:

$$f(x_1(t), x_2(t), x_3(t)) = f_1(x_1(t)) f_2(x_2(t)) f_3(x_3(t)) \cdot c_{1,2}(F_1(x_1(t)), F_2(x_2(t))) c_{2,3}(F_2(x_2(t)), F_3(x_3(t))) \cdot c_{1,3|2}(F_{1|2}(x_1(t)|x_2(t)), F_{3|2}(x_3(t)|x_2(t))).$$



- Joint modeling of y and its covariates $x_1, \ldots, x_p$ by D-vine copula based quantile regression (DVQR), i.e.

$$F_{y|x_1,\ldots,x_p}^{-1}(\alpha|x_1(t), \ldots, x_p(t)) := F_y^{-1}\left(C^{-1}(\alpha|F_{x_1}(x_1(t)), \ldots, F_{x_p}(x_p(t)))\right),$$

where C^{-1} denotes the inverse conditional copula quantile function obtained from the D-vine copula.

Model estimation

- Marginal functions $F_\bullet$ are estimated via kernel density estimates.
- Large set of copula families available (elliptical, archimedean, nonparametric copulas).
- Automated forward predictor selection by minimizing a conditional BIC.
- Variable importance criteria for a $(p+1)$-dimensional D-vine copula with $x_1, \ldots, x_m$ predictors and order $y - x_{j_1} - \ldots - x_{j_p}$:

$$\text{Imp}(x_j) := \begin{cases} 1 - \dfrac{s_j - 1}{m}, & x_j \in \{x_{j_1}, \ldots, x_{j_p}\}, \\ 0, & \text{otherwise}, \end{cases}$$

where s_j denotes the position of x_j in $x_{j_1} - \ldots - x_{j_p}$.

5. Setting

Local PP

Observation and forecast data from all available stations are combined to estimate a single model valid for all stations.

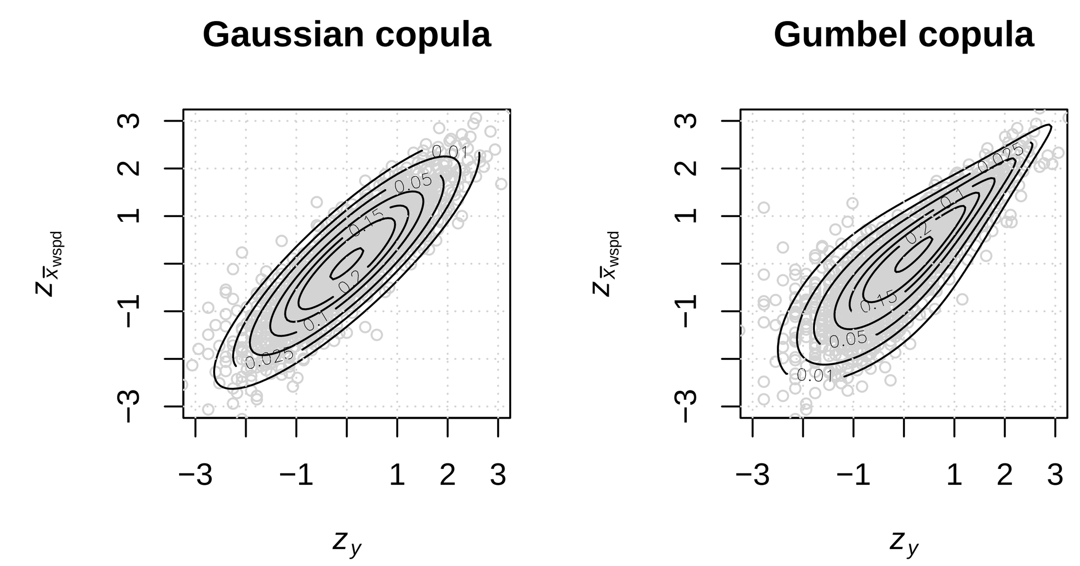
- Training: Allow all copula families and the following predictors:

Method	Predictors $\overline{x}_\bullet, x_\bullet^{\text{ctrl}}$ of ...
DVQR$^-$	wspd
DVQR$^+$	wspd, u, v, angle, wgust, temp, sp, tcc, sh

- Validation: Reduce the copula family set for all stations and select the predictors and its order according to the variable importance criteria (Imp) for each station separately.

Method	Amount of predictors	Amount of copula families
DVQR$^-$	1 (mostly $\overline{x}_{\text{wspd}}$)	2 (Gaussian, Gumbel)
DVQR$^+$	5 (mostly $\overline{x}_{\text{wspd}}, \overline{x}_{\text{wugst}}, \overline{x}_{\text{u}}, \overline{x}_{\text{v}}, \overline{x}_{\text{tcc}}$)	6

- Empirical pairwise contour (dependence) plots for y and $\overline{x}_{\text{wspd}}$:



Global PP

Observation and forecast data from all available stations are combined to estimate a single model valid for all stations.

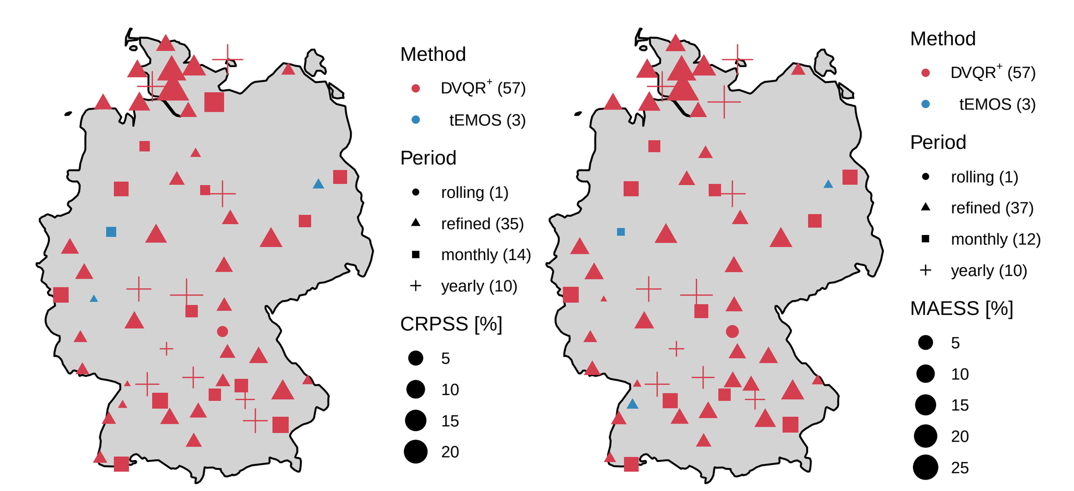
- Training: Allow all copula families and the same predictor variables as for the local PP except for DVQR$^+$, where we add $x_{\text{lon}}, x_{\text{lat}}, x_{\text{elev}}$.
- Validation: Reduce the copula family set and select the predictors and its order according to the variable importance criteria (Imp) for all stations.

Method	Amount of predictors	Amount of copula families
DVQR$^-$	2 with fixed order $\overline{x}_{\text{wspd}} - x_{\text{wspd}}^{\text{ctrl}}$	3 (Gumbel, T, TLL)
DVQR$^+$	7 (mostly $\overline{x}_{\text{wspd}}, \overline{x}_{\text{wugst}}, \overline{x}_{\text{tcc}}, \overline{x}_{\text{elev}}, \overline{x}_{\text{lat}}$)	7

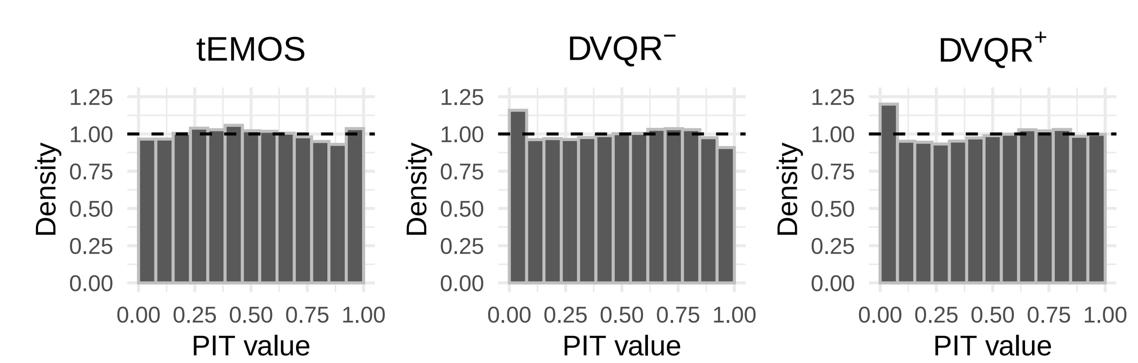
6. Results

Local PP

- DVQR$^-$, DVQR$^+$ and tEMOS improve the raw ensemble about 28%-31% with respect to CRPS.
- DVQR$^+$ outperforms tEMOS up to 6% with respect to CRPS.
- Best method and training period based on mean CRPS (left) or MAE (right) incl. skill score improvement to the competing method:

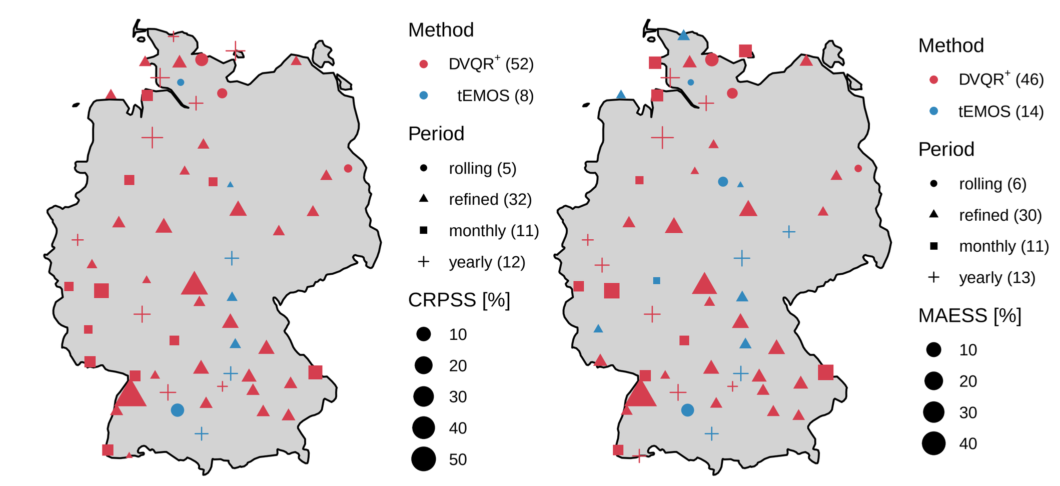


- DVQR$^+$ yields sharper forecasts than tEMOS, but tEMOS yields slightly more calibrated forecasts than DVQR$^+$.
- PIT histograms aggregated over all validation days, stations and types of training periods:

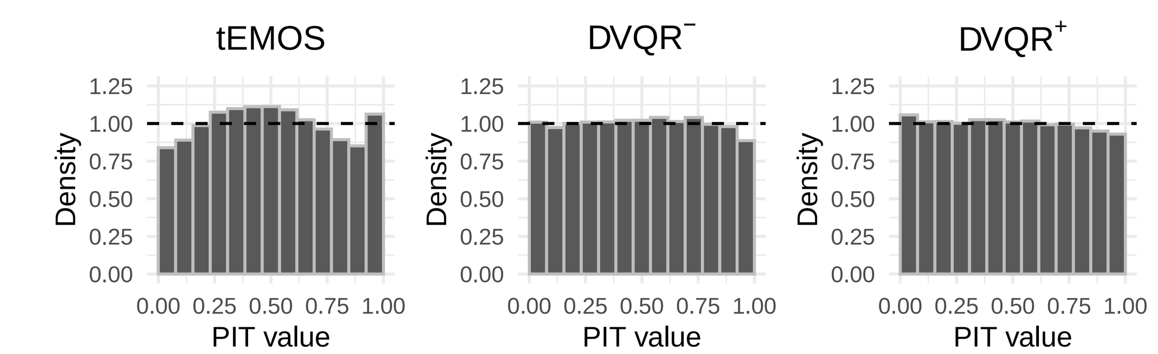


Global PP

- DVQR$^+$ outperforms tEMOS up to 8% with respect to CRPS.
- Best method and training period based on mean CRPS (left) or MAE (right) incl. skill score improvement to the competing method:

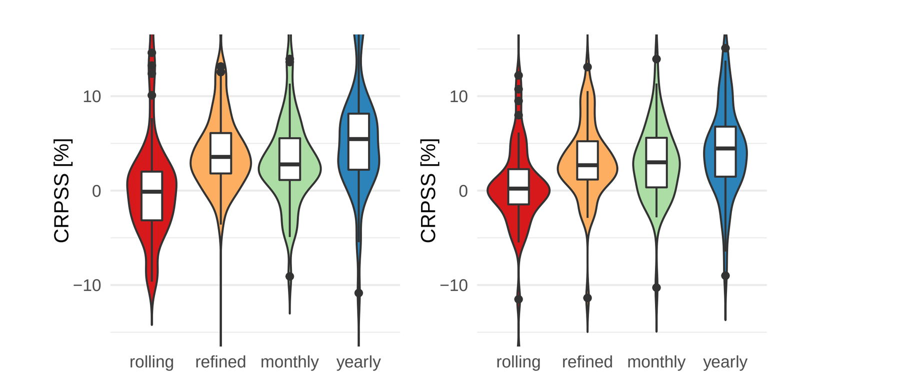


- tEMOS produces in overall sharper forecasts than DVQR$^+$, while DVQR$^+$ yields more calibrated forecasts.
- PIT histograms aggregated over all validation days, stations and types of training periods:



General results

- Local DVQR and tEMOS models yield better results (with respect to e.g. CRPS, MAE) than their global variants.
- No strong performance differences between DVQR$^-$ and tEMOS for local & global PP.
- The refined training period performs the best with respect to CRPS followed by the monthly training period.
- CRPS skill score improvement of DVQR$^+$ over tEMOS for the local (left) and global (right) models depending on the type of training period:



7. Conclusion

- DVQR identifies the most important predictor variables.
- wspd, wgust, u, v, tcc are promising predictor variables for wind speed PP.
- DVQR is able to significantly outperform tEMOS.
- Refined and monthly training period are very suitable for tEMOS and DVQR wind speed PP.

References

David Jobst, Annette Möller, Jürgen Groß. *D-Vine Copula based Postprocessing of Wind Speed Ensemble Forecasts.* Quaterly Journal of the Royal Meteorological Society. (under review)

Correspondence: David Jobst, University of Hildesheim, jobst@imai.uni-hildesheim.de