Vine copula application to postprocessing

Natalia Szopa, Institute of Meteorology and Water Management

Introduction

- Work supervised by Bogdan Bochenek of IMGW-PIB, Joanna Czarnowska of University of Gdańsk and Dawid Tarłowski of Jagiellonian University.
- Results have been presented in a Mathematics master's degree dissertation.
- Aim of the presentation: to show a statistical post processing method based on vine copulas which could be used for the correction of systematic errors of the NWP model.

Motivation for the copula based method

- Flexible modeling of dependencies structure between multiple variables.
- Ability to draw samples from conditional distributions depending on different choice conditional variables.
- Generation of synthetic data that models pre-existing multivariate data.

- Core of our method:
  - Goal: Estimating the error of 2m air temperature forecast of the ALARO model.
  - Can we provide a correction of the forecast by using a generated sample from a conditional copula probability distribution?

Mathematical background of the copula approach

- A *d*-dimensional **copula** $\boldsymbol{C}(\boldsymbol{x_1}, \dots, \boldsymbol{x_d})$ is a multivariate distribution function on $[0,1]^d$ with **uniformly distributed marginals**.
- **Sklar's Theorem:**

  For a d-dimensional cumulative distribution function, there exists a copula $\boldsymbol{C}$, such that $\boldsymbol{F}\left(\boldsymbol{x_1}(\boldsymbol{t}), .., \boldsymbol{x_d}(\boldsymbol{t})\right) = \mathbf{C}\left(\boldsymbol{F_1}(\boldsymbol{x_1}(\boldsymbol{t})), \dots, \boldsymbol{F_d}\left(\boldsymbol{x_d}(\boldsymbol{t})\right)\right)$, where $\boldsymbol{F}$ is a joint cumulative distribution function and $\boldsymbol{F_1}, \dots, \boldsymbol{F_d}$ are marginal distribution functions.

  - This theorem allows to separate univariate margins from the dependence structure.
  - Easy way of constructing a wide range of more flexible multivariate distributions.

Pair copula decompositions and constructions

- A way to construct multivariate copulas using only bivariate copulas as building blocks done by recursive factorization.
- For example we can express an arbitrary joint three dimensional probability density in terms of marginal densities, bivariate copula densities and conditional distribution functions as follows:

$$f(x_1,x_2,x_3) = c_{13;2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2); x_2) \times c_{23}(F_2(x_2), F_3(x_3)) \\ \times c_{12}(F_1(x_1), F_2(x_2)) f_3(x_3) f_2(x_2) f_1(x_1).$$

- This decomposition is not unique since we can express it as:

$$f(x_1,x_2,x_3) = c_{23;1}(F_{2|1}(x_2|x_1), F_{3|1}(x_3|x_1); x_1) \times c_{13}(F_1(x_1), F_3(x_3)) \\ \times c_{12}(F_1(x_1), F_2(x_2)) f_3(x_3) f_2(x_2) f_1(x_1)$$

or

$$f(x_1,x_2,x_3) = c_{12;3}(F_{1|3}(x_1|x_3), F_{2|1}(x_2|x_1); x_3) \times c_{13}(F_1(x_1), F_3(x_3)) \\ \times c_{23}(F_2(x_2), F_3(x_3)) f_3(x_3) f_2(x_2) f_1(x_1)$$

- Key takeaway: Different decompositions depend on the choice and order of conditioning variables.
- Special cases of decomposition: C-vines and D-vines which are possible to be presented as graphs called vine tree structures.

Data

- The data we used in the study include forecasts from three numerical weather models: ALARO (res. 4 x 4 km), AROME (res. 2 x 2 km) and COSMO (res. 7 x 7 km) for 35 Polish meteorological stations in the years 2019 and 2020.
- Training set:
  - forecasts from 01.01.2019 – 31.12.2019
- Test set:
  - forecasts from 01.01.2020 – 31.12.2020

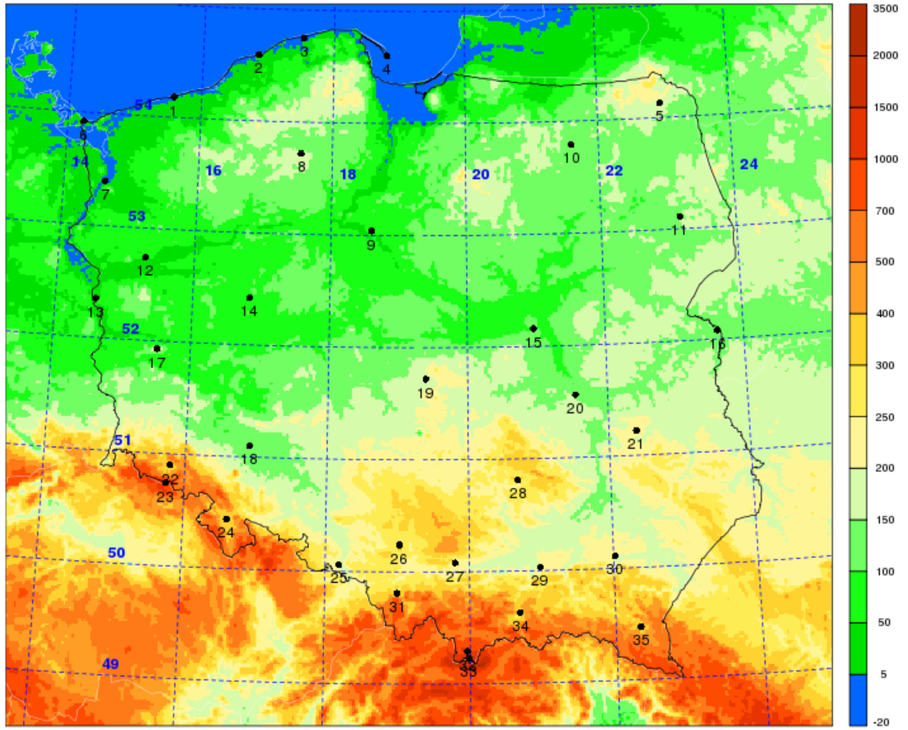

Figure 1. Orography map of the study area with country borders (black line) and locations of synoptic stations (black dots with station ID described below).

Method

- Goal: **error mitigation** of the 2m above ground level air temperature forecast given by the ALARO model.
- Outline:

1. Fitting a **vine copula model** which best describes the dependency structure between the variables affecting the forecast error and their individual probability distributions.
2. Obtaining a sample of pseudo-observations from a copula-given conditional probability distribution of the **ALARO model error** with different **conditioning variables** (described on the next slide).
3. Checking whether the choice of different **conditioning variables** has a significant effect on the correct fit of the model.

Method

Goal: **error mitigation** of the air temperature 2m above ground level forecast given by the ALARO model.

1. Fitting either a C- or D-vine copula model.
2. Drawing a **10000-element sample** from the selected copula model.
3. **Adding the mean** of the generated forecast errors to the temperature forecast of the ALARO model.
4. Checking the accuracy of the correction using RMSE and bias.

Description of the conditioning variables	Indicator
AROME model forecast for the current day	a
COSMO model forecast for the current day	b
Forecast error of the ALARO model on the previous day	c
Value of observed temperature at 00 UTC	d
Forecast error of the AROME model on the previous day	e
Forecast error of the COSMO model on the previous day	f
Forecast error of the AROME model on the current day	g
Forecast error of the COSMO model on the current day	h
Difference between the forecast on the previous day and the current day of the ALARO model	i
Difference between the previous day's relative humidity forecast and the current day's ALARO model forecast	j

Verification

- The effectiveness of the method was evaluated using the root mean square error (RMSE) and bias.

$$\mathrm{RMSE}[^\circ\mathrm{C}] = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2}$$

$$\mathrm{bias}[^\circ\mathrm{C}] = \frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)$$

- The predicted values in the 2019 test set were verified against the measured temperature values in the 2020 training set.

Results

- Significant reductions in RMSE were observed at Śnieżka (Station: 23), in Hel (4), in Zakopane (32) and on Kasprowy Wierch (33)
- The greatest reduction in error is seen at the stations where this error was originally greatest.
- On average, the largest improvement observed with the three conditioning variables: 2 m air temperature forecasts of AROME model *(a)* and COSMO model *(b)* forecasts initialized at 00 UTC with lead time 12h and the value of the observed 2 m air temperature at 00 UTC *(d)*

Table 2. RMSE of the ALARO T2m forecast depending on different sets of conditioning variables.

Station no.	RMSE for 2020	a, b, c	a, b, d	e, f	e, f, i	e, f, j	Average RMSE after correction	Average percentage change in RMSE
1	1.571	1.567	1.444	1.459	1.445	1.432	1.469	6%
2	1.319	1.366	1.283	1.381	1.279	1.367	1.335	-1%
3	1.365	1.364	1.348	1.394	1.394	1.381	1.376	-1%
4	2.317	1.828	1.837	1.536	1.450	1.443	1.619	28%
5	1.397	1.454	1.255	1.351	1.517	1.410	1.398	0%
6	1.522	1.305	1.318	1.354	1.325	1.335	1.327	13%
7	1.538	1.352	1.329	1.445	1.338	1.613	1.415	11%
8	1.375	1.373	1.337	1.333	1.309	1.311	1.333	3%
9	1.568	1.509	1.459	1.550	1.536	1.506	1.512	3%
10	1.591	1.626	1.473	1.536	1.433	1.407	1.495	5%
11	1.496	1.649	1.412	1.716	1.587	1.617	1.596	-6%
12	1.408	1.358	1.367	1.346	1.364	1.315	1.350	4%
13	1.743	1.662	1.568	1.644	1.630	1.655	1.632	7%
14	1.459	1.432	1.459	1.510	1.491	1.487	1.476	-1%
15	1.454	1.407	1.342	1.441	1.441	1.543	1.435	3%
16	1.553	1.485	1.469	1.565	1.501	1.621	1.528	3%
17	1.573	1.375	1.390	1.392	1.425	1.367	1.390	11%
18	1.408	1.359	1.362	1.436	1.411	1.413	1.396	1%
19	1.402	1.384	1.340	1.412	1.367	1.357	1.372	2%
20	1.642	1.684	1.669	1.778	1.754	1.692	1.715	-5%
21	1.362	1.403	1.432	1.418	1.434	1.397	1.417	-4%
22	1.626	1.759	1.510	1.589	1.843	1.967	1.734	-3%
23	3.738	2.015	2.279	2.006	1.885	1.880	2.013	45%
24	1.596	1.569	1.465	1.577	1.520	1.438	1.514	4%
25	1.526	1.496	1.424	1.456	1.457	1.413	1.449	4%
26	1.507	1.444	1.445	1.450	1.508	1.568	1.483	3%
27	1.697	1.753	1.720	1.723	1.846	1.789	1.766	-4%
28	1.450	1.577	1.493	1.527	1.563	1.599	1.552	-6%
29	1.759	1.833	1.760	1.810	1.864	1.823	1.818	-3%
30	1.542	1.562	1.540	1.576	1.525	1.563	1.553	-1%
31	1.810	1.791	1.736	1.816	1.880	1.827	1.810	0%
32	3.479	2.833	2.842	2.706	2.962	2.289	2.726	18%
33	2.949	2.285	2.542	2.480	2.430	2.434	2.434	17%
34	1.771	1.896	1.789	1.972	1.937	1.811	1.881	-7%
35	1.695	1.907	1.685	1.809	1.769	1.646	1.763	-6%
Average	**1.618**	**1.509**	**1.462**	**1.503**	**1.486**	**1.498**		

Results

- Insignificant reduction of the bias of the ALARO model.
- Underestimation of the forecast both before and after applying the error correction.
- The best-fitting copula was the one conditioned

on the:

  - Forecast error of the AROME model on the previous day *(e)*
  - Forecast error of the COSMO model on the previous day *(f)*
  - Difference between the forecast on the previous day and the current day of the ALARO model *(i)*

Table 3. Mean bias of the ALARO T2m forecast depending on different sets of conditioning variables.

Station	Mean bias for 2020	*a, b, c*	*a, b, d*	*e, f*	*e, f, i*	*e, f, j*	Average bias after correction
1	-0.386	-0.314	-0.336	-0.251	-0.142	-0.191	-0.247
2	-0.152	-0.256	-0.181	-0.067	-0.134	-0.156	-0.159
3	0.036	-0.195	-0.147	-0.162	-0.16	-0.156	-0.164
4	1.28	-0.113	0.193	-0.13	-0.105	-0.111	-0.053
5	-0.279	-0.256	-0.093	-0.111	-0.061	-0.052	-0.115
6	0.741	-0.088	0.069	-0.025	0.008	0.014	-0.005
7	-0.315	0.075	-0.022	-0.042	0.049	0.172	0.046
8	-0.354	-0.106	-0.228	-0.077	-0.079	-0.086	-0.115
9	-0.124	-0.062	-0.075	-0.072	0.05	0.012	-0.029
10	-0.661	-0.604	-0.57	-0.529	-0.292	-0.279	-0.455
11	-0.291	-0.58	-0.312	-0.373	-0.139	-0.209	-0.323
12	-0.282	0.036	-0.117	0.014	0.042	0.081	0.011
13	-0.594	-0.252	-0.235	-0.135	-0.106	0.125	-0.121
14	-0.223	-0.234	-0.267	-0.157	-0.115	-0.056	-0.166
15	-0.337	-0.371	-0.301	-0.463	-0.249	-0.077	-0.292
16	-0.438	-0.601	-0.488	-0.528	-0.257	-0.531	-0.481
17	-0.722	-0.105	-0.222	-0.169	-0.03	-0.037	-0.113
18	0.077	-0.06	0.065	-0.005	-0.009	-0.019	-0.006
19	-0.286	-0.308	-0.325	-0.31	-0.144	-0.065	-0.23
20	-0.28	-0.257	-0.499	-0.49	-0.217	-0.057	-0.304
21	-0.271	-0.389	-0.358	-0.296	-0.159	-0.196	-0.28
22	0.485	-0.733	-0.261	-0.396	-0.588	-0.444	-0.484
23	-3.078	0.493	0.254	0.49	0.09	0.075	0.281
24	-0.617	-0.513	-0.299	-0.341	-0.322	-0.274	-0.35
25	-0.31	0.171	0.189	0.106	0.167	0.131	0.153
26	0.223	-0.132	0.02	-0.148	-0.103	-0.089	-0.091
27	0.046	-0.629	-0.458	-0.503	-0.581	-0.466	-0.527
28	-0.007	-0.406	-0.307	-0.347	-0.283	-0.274	-0.323
29	0.227	-0.583	-0.439	-0.505	-0.47	-0.409	-0.481
30	-0.147	-0.258	-0.121	-0.261	-0.12	-0.125	-0.177
31	0.191	-0.308	-0.195	-0.184	-0.123	-0.388	-0.24
32	2.114	0.073	-0.249	0.064	-0.027	-0.196	-0.067
33	-1.644	0.642	0.782	0.538	0.142	0.446	0.51
34	0.359	-0.712	-0.497	-0.7	-0.521	-0.498	-0.585
35	-0.168	-0.486	-0.242	-0.285	-0.414	-0.261	-0.338
Average	**-0.177**	**-0.241**	**-0.179**	**-0.196**	**-0.148**	**-0.083**	

Conclusions

- A slight correction in the temperature prediction of the ALARO model is noted.
- Greatest improvement is seen at the “most outlying” meteorological stations, such as:
  - the top of Śnieżka (1613m above sea level) and Kasprowy Wierch (1989m above sea level), Zakopane (857m above sea level),
  - stations located close to the seaside such as Hel, Świnoujście and Szczecin.
- Time needed for computing is a big disadvantage - over 2 hours to fit a right copula distribution and then estimate the error (the tests have been conducted on 1 node with 16 cores (each core 128GB).
- In the future the script could be improved for enhancing the efficiency.
- Copulas might be more useful in other applications such as analyzing the dependence structure between weather elements in compound events such as floods, as described in Bevacqua et al. (2017).

References

- Bevacqua, Emanuele & Maraun, Douglas & Hobaek Haff, Ingrid & Widmann, Martin & Vrac, Mathieu. (2017). Multivariate statistical modelling of compound events via pair-copula constructions: Analysis of floods in Ravenna (Italy). Hydrology and Earth System Sciences. 21. 2701-2723. 10.5194/hess-21-2701-2017.

- Bevacqua E (2017). CDVineCopulaConditional: Sampling from Conditional C- and D-Vine Copulas. R package version 0.1.0, https://CRAN.R-project.org/package=CDVineCopulaConditional.

- Czado C (2019) Analyzing Dependent Data with Vine Copulas: a Practical Guide with R, vol 222, 1st edn. Lecture Notes in Statistics. Springer, Cham, Switzerland

- Nagler, T., Schellhase, C., and Czado, C. (2017). Nonparametric estimation of simplified vine copula models: comparison of methods. Dependence Modeling, 5:99-120.

- Perrone, Elisa & Schicker, Irene & Lang, Moritz. (2020). A case study of empirical copula methods for the statistical correction of forecasts of the ALADIN-LAEF system. Meteorologische Zeitschrift. 29. 10.1127/metz/2020/1034.