QUANTIFYING UNCERTAINTIES IN OBSERVATIONAL DATASETS OVER THE CARPATHIAN REGION

Tímea Kalmár, Erzsébet Kristóf, Rita Pongrácz, Ildikó Pieczka and Roland Hollós
Department of Meteorology, Eötvös Loránd University, Budapest, Hungary

EGU22
27th May, 2022

ELTE
EÖTVÖS LORÁND
UNIVERSITY

This presentation participates in OSPP
EGU
Outstanding Student & PhD candidate Presentation contest

Motivation and objective

Gridded observational datasets are essential for the evaluation of climate simulations.

However, uncertainty originating from the selection of observations is as important as the uncertainty of regional climate models

In this study, a novel evaluation method is introduced which combines widely known metrics and statistical techniques (e.g., comparison of spatio-temporal distributions by relative difference ($DIFF_{rel}$), root mean square error (RMSE), temporal Pearson correlation coefficient (r_t), applying k-means clustering) to quantify uncertainties in the observational datasets

1. Comparing temporal distributions of daily precipitation obtained from the observational datasets

1. Comparing temporal distributions of daily precipitation obtained from the observational datasets

2. Comparing spatial distributions of variables in each observational dataset:
Annual sum of precipitation (PR)
Elevation (E)
Variability of elevation (**VE**)
Effect of station density (**PR_ST**)

- **VE**: we apply a moving window filter to the data, then we compute the difference between the largest and the lowest elevation values within the window and the difference is assigned to the central cell of the window.
- **PR_ST**: we count how many stations are located within the window, and the number is assigned to the central cell of the window.

1. Comparing temporal distributions of daily precipitation obtained from the observational datasets

2. Comparing spatial distributions of variables in each observational dataset:
Annual sum of precipitation (PR)
Elevation (E)
Variability of elevation (**VE**)
Effect of station density (**PR_ST**)

1. Spatial correlation coefficient (r_s) between all possible pairs of variables
2. Random sampling of 5895 grid cells and permutation test to determine the significance of r_s → 10,000 original r_s ($r_{s,original}$) and random r_s ($r_{s,random}$)
3. Calculating uncertainty (U) to determine the relaibility of the pairs of variables: overlapping area of the probability density functions (PDFs) fitted on the histograms of the $r_{s,original}$ and $r_{s,random}$ values.
U increases with increasing overlapping area that means less reliable linear relationship between the variables.
4. To distinguish the pairs of variables objectively with respect to their reliabilites, k-mean clustering was applied on their median of $r_{s,original}$ values (r) and U values

- **VE**: we apply a moving window filter to the data, then we compute the difference between the largest and the lowest elevation values within the window and the difference is assigned to the central cell of the window.
- **PR_ST**: we count how many stations are located within the window, and the number is assigned to the central cell of the window.

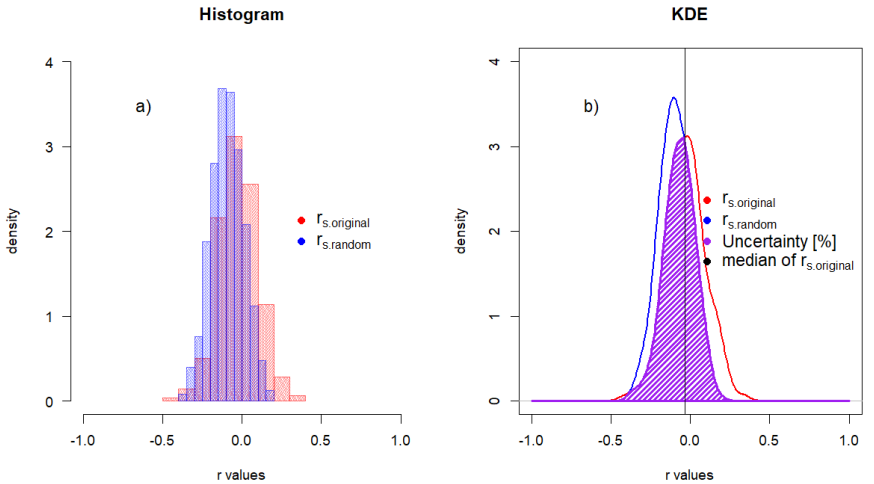

Figure I Illustration of probability density estimation

Datasets

- CarpatClim and E-OBSv22
- Domain: 17-27°E; 44-50°N
- Horizontal resolution: ~10 km (0.1°)
- Time period: 2010

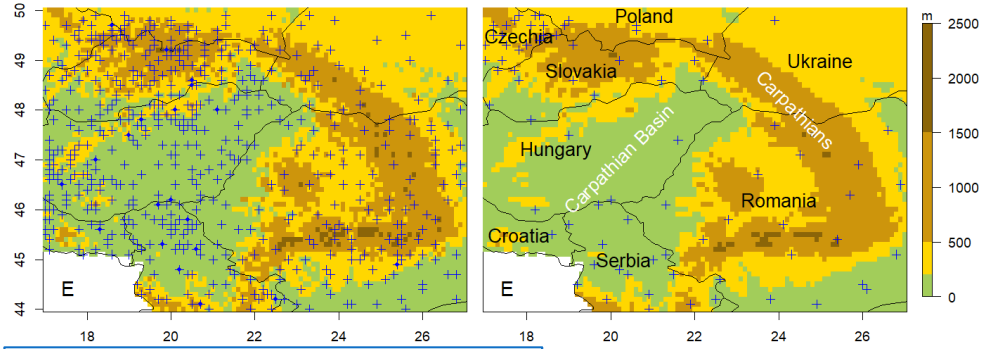

Figure 2 The location of precipitation stations (indicated by blue +, if a grid cell contains two stations, it is marked by blue dot) for CarpatClim (left) and E-OBS (right).

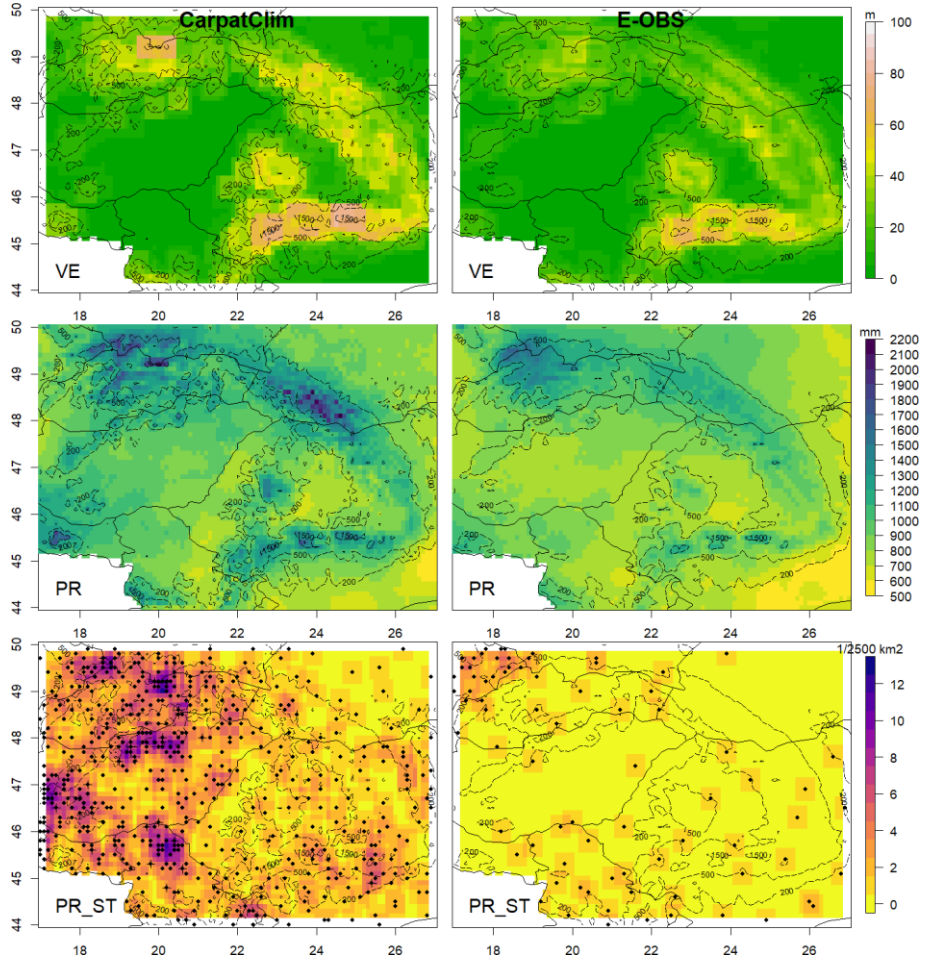

Figure 3 Spatial distribution of VE, PR and PR_ST for CarpatClim (left) and E-OBS (right)

Temporal distribution: CarpatClim vs E-OBS

- $DIFF_{rel}$: EOBS underestimation the precipitation by 50%
- RMSE: low station density over Northeastern Carpathians causes high RMSE values
- r_t: shifting in daily precipitation causes the problem in Serbia → correction
- Around the E-OBS stations, the RMSE values are close to 0 and r_t values are close to 1 (interpolation)

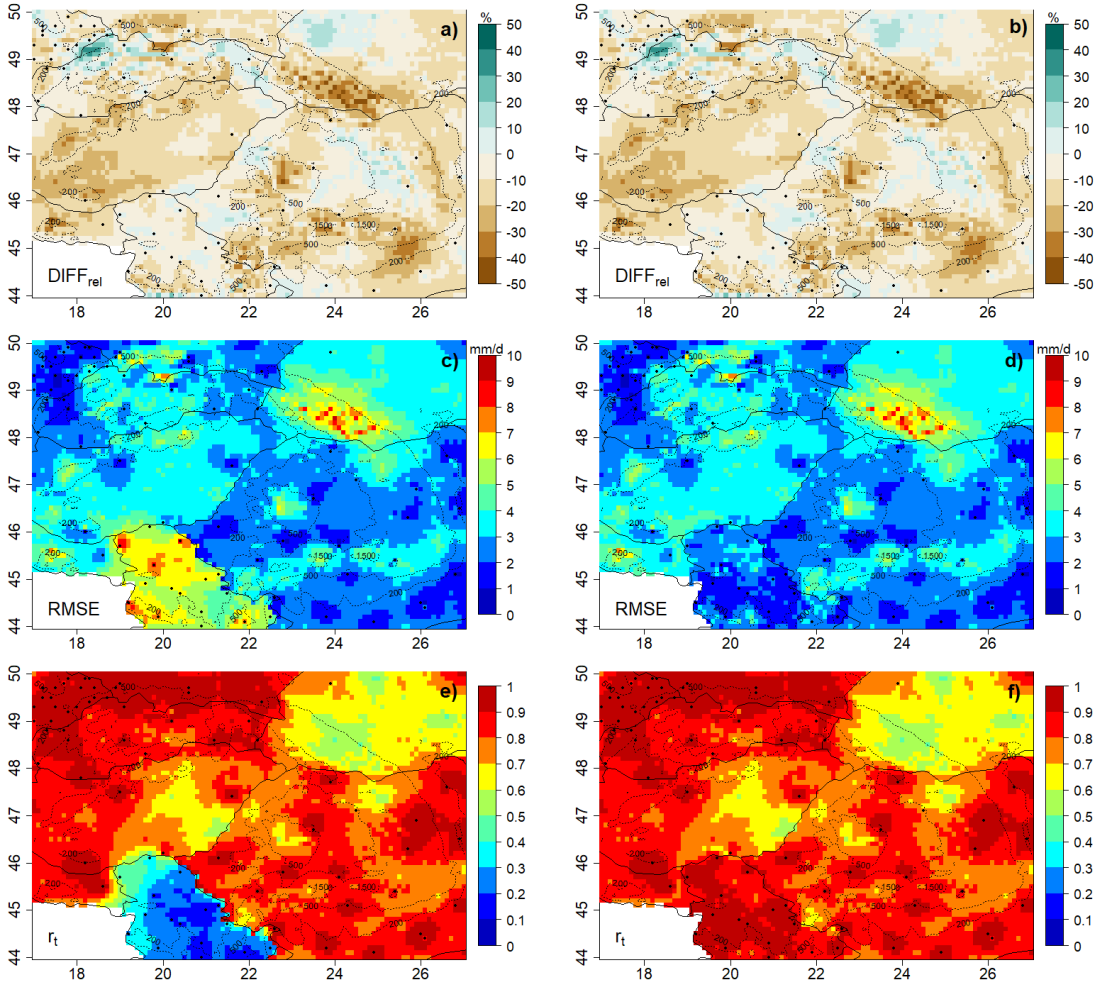

Figure 4 Relative difference ($DIFF_{rel}$, a-b), RMSE (c-d) and temporal correlation (r_t, e-f) between CarpatClim and E-OBS. Maps of the left (right) column is based on the stations from E-OBS before (after) the database correction. The black dots indicate the precipitation stations from E-OBS dataset.

Spatial distribution of the variables

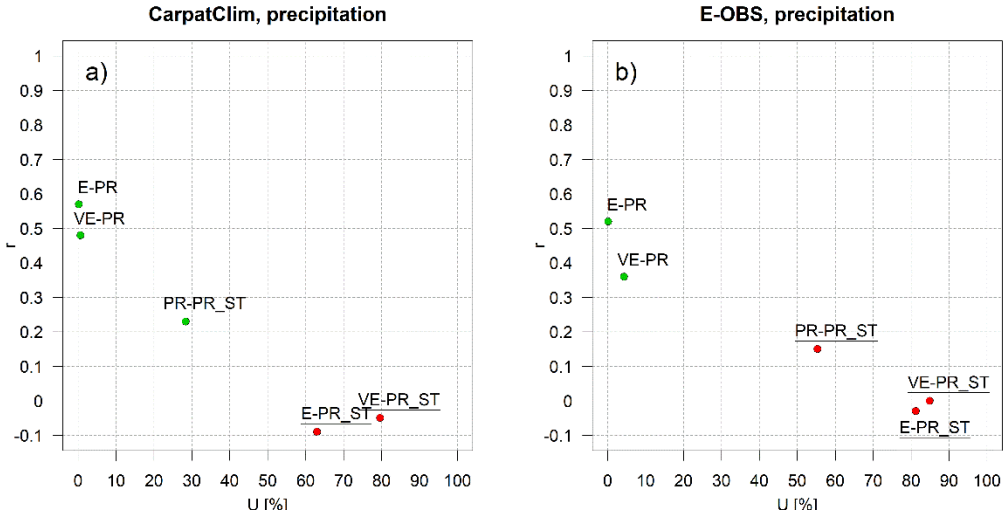

Figure 4 Scatterplots for uncertainty (U) and the median of original correlation coefficients (r) obtained from CarpatClim (left) and E-OBS (right) for precipitation. Two clusters of the pairs of variables are: the first cluster (which contains pairs of variables with reliable relationship) is distinguished from the second cluster (which contains pairs of variables with less reliable relationship) with underlined variables. The green (red) dots show the pairs of variables with significant (non-significant) r values at a level of 0.05.

- E-PR and EV-PR relationships are stronger in CarpatClim than in E-OBS
- PR_ST is significant only with PR in case of CarpatClim
- Two clusters are detected: the first cluster contains pairs of variables, which relationships are considered reliable, i.e., strong correlations ($r>0.4$ and $r<-0.4$) and small U values ($U<30\%$). The second cluster contains pairs of variables, which relationships are considered less reliable, i.e., weak correlations ($-0.2<r<0.2$) and large U values ($U>30\%$).

Spatial distribution of the variables

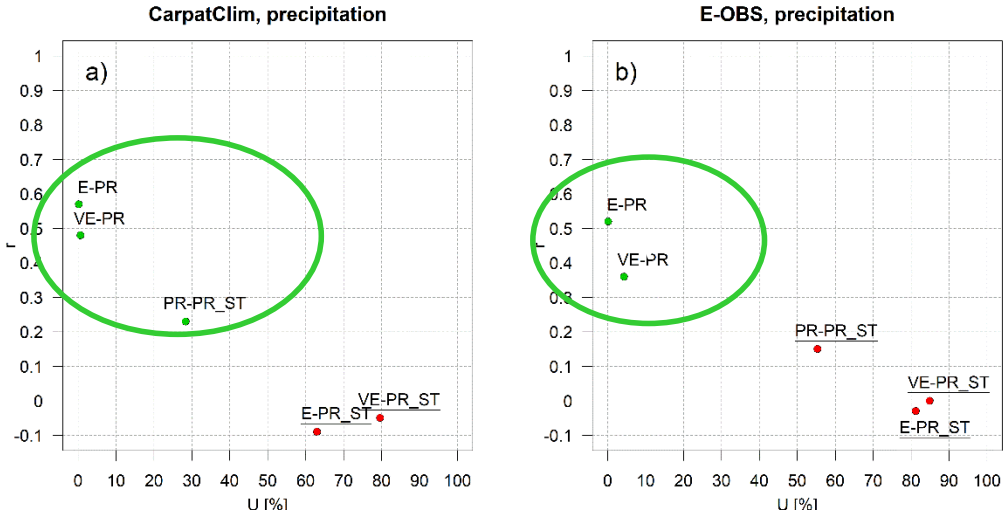

Figure 4 Scatterplots for uncertainty (U) and the median of original correlation coefficients (r) obtained from CarpatClim (left) and E-OBS (right) for precipitation. Two clusters of the pairs of variables are: the first cluster (which contains pairs of variables with reliable relationship) is distinguished from the second cluster (which contains pairs of variables with less reliable relationship) with underlined variables. The green (red) dots show the pairs of variables with significant (non-significant) r values at a level of 0.05.

- E-PR and EV-PR relationships are stronger in CarpatClim than in E-OBS
- PR_ST is significant only with PR in case of CarpatClim
- Two clusters are detected: the first cluster contains pairs of variables, which relationships are considered reliable, i.e., strong correlations ($r>0.4$ and $r<-0.4$) and small U values ($U<30\%$). The second cluster contains pairs of variables, which relationships are considered less reliable, i.e., weak correlations ($-0.2<r<0.2$) and large U values ($U>30\%$).

Spatial distribution of the variables

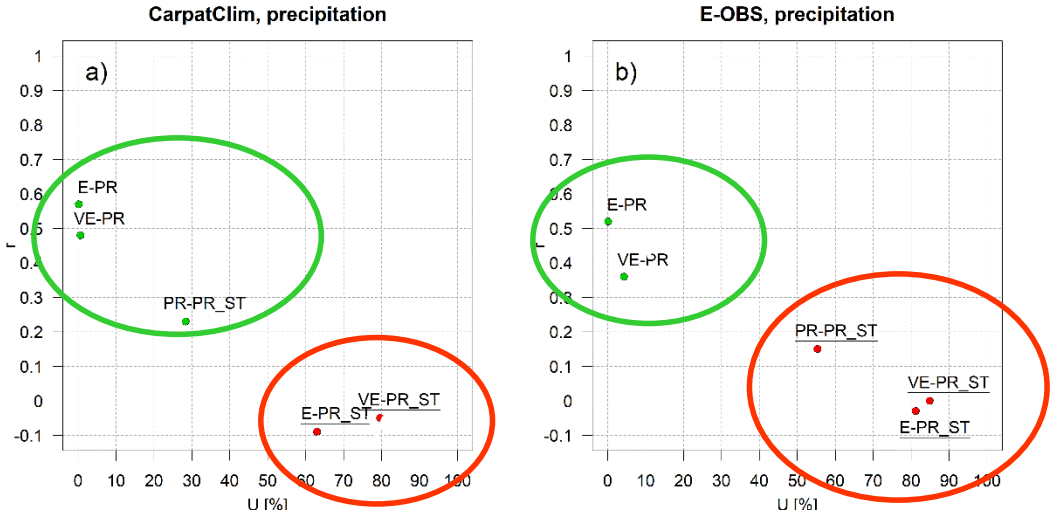

Figure 4 Scatterplots for uncertainty (U) and the median of original correlation coefficients (r) obtained from CarpatClim (left) and E-OBS (right) for precipitation. Two clusters of the pairs of variables are: the first cluster (which contains pairs of variables with reliable relationship) is distinguished from the second cluster (which contains pairs of variables with less reliable relationship) with underlined variables. The green (red) dots show the pairs of variables with significant (non-significant) r values at a level of 0.05.

- E-PR and EV-PR relationships are stronger in CarpatClim than in E-OBS
- PR_ST is significant only with PR in case of CarpatClim
- Two clusters are detected: the first cluster contains pairs of variables, which relationships are considered reliable, i.e., strong correlations ($r>0.4$ and $r<-0.4$) and small U values ($U<30\%$). The second cluster contains pairs of variables, which relationships are considered less reliable, i.e., weak correlations ($-0.2<r<0.2$) and large U values ($U>30\%$).

Conclusions

- Through our comprehensive analysis, we pointed out that the analysis of the temporal distribution of the variables is useful for error detection
- The spatial distribution of the examined climatic and geographical variables shows that CarpatClim is wetter over the whole region (mostly over the mountains) than E-OBS
- The joint investigation of spatial correlations between the pairs of variables and the associated uncertainties were useful to distinguish the pairs of variables based on reliability of their relationships
- CarpatClim is more reliable compared to E-OBS in case of precipitation
- The method was applied to CarpatClim and E-OBS, but it could be applied to other datasets, different time periods and geographical areas.

Conclusions

- Through our comprehensive analysis, we pointed out that the analysis of the temporal distribution of the variables is useful for error detection
- The spatial distribution of the examined climatic and geographical variables shows that CarpatClim is wetter over the whole region (mostly over the mountains) than E-OBS
- The joint investigation of spatial correlations between the pairs of variables and the associated uncertainties were useful to distinguish the pairs of variables based on reliability of their relationships
- CarpatClim is more reliable compared to E-OBS in case of precipitation
- The method was applied to CarpatClim and E-OBS, but it could be applied to other datasets, different time periods and geographical areas.

Thanks for your attention!

E-mail: timea.kalmar@ttk.elte.hu