DTU Library

Comparison of data-driven methods for linking extreme precipitation events to large-scale drivers

A case study from Copenhagen, Denmark

Antoniadou, Nafsika; Sørup, Hjalte Jomo Danielsen; Pedersen, Jonas Wied; Gregersen, Ida Bülow; Schmith, Torben; Arnbjerg-Nielsen, Karsten

Link to article, DOI:
10.5194/egusphere-egu23-5074

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

Citation (APA):
Antoniadou, N., Sørup, H. J. D., Pedersen, J. W., Gregersen, I. B., Schmith, T., & Arnbjerg-Nielsen, K. (2023). *Comparison of data-driven methods for linking extreme precipitation events to large-scale drivers: A case study from Copenhagen, Denmark*. Abstract from EGU General Assembly 2023, Vienna, Austria. https://doi.org/10.5194/egusphere-egu23-5074

EGU23-5074, updated on 09 May 2023
https://doi.org/10.5194/egusphere-egu23-5074
EGU General Assembly 2023

Comparison of data-driven methods for linking extreme precipitation events to large-scale drivers: A case study from Copenhagen, Denmark

Nafsika Antoniadou[1,2], Hjalte Jomo Danielsen Sørup[1], Jonas Wied Pedersen[1,2], Ida Bülow Gregersen[3], Torben Schmith[2], and Karsten Arnbjerg-Nielsen[1]

[1]Department of Environmental and Resource Engineering, Technical University of Denmark, Kgs. Lyngby, Denmark
[2]Danish Meteorological Institute, Copenhagen, Denmark
[3]Ramboll DK, Copenhagen, Denmark

Extreme precipitation events can lead to severe negative consequences on society, the economy, and the environment. To mitigate related risks, it is crucial to understand their natural causes. There is a vast number of methods in the literature analyzing their connection to large-scale drivers. Recently there has been much interest in using machine learning (ML) methods instead of traditional statistical models like regression. ML methods are based on algorithms adapting and learning from data. By contrast, regression models are based on theory and assumptions and benefit from domain knowledge for model specification. Because of its adaptability, ML is claimed to offer superior predictive performance than traditional statistical modeling and better manage a greater number of potential predictors. A few studies in climate research have compared the performance between these two approaches, but their conclusions are inconsistent, and some have limitations.

We used five predictor variables - Geopotential height at 500hPA, Convective available energy (CAPE), Total column water (TCW), Sea Surface Temperature (SST), and Surface Temperature (SAT) using ERA5, the latest reanalysis dataset from ECMWF, and data produced by the Danish Meteorological Institute. All the predictors were not used directly as inputs but were preprocessed before modeling. We trained models using logistic regression (LR) and three commonly used supervised machine learning algorithms - random forests (RF), neural networks (NNET), and support vector machines (SVM) to predict whether an extreme event occurred over Copenhagen. In the LR framework, the predictor variables were modeled using restricted cubic splines to address potential nonlinearity. The training data are highly unbalanced, so using a traditional performance metric such as accuracy (ACC) could be misleading. In light of this, we use performance metrics specialized for unbalanced datasets: the ROC (receiver operating characteristic) curve as the primary measure and the area under the precision-recall curve, the Brier score, and ACC together with the true positive rate and the false positive rate at the optimal threshold as secondary measures.

During the variable selection process, it was found that SST has the weakest relationship with

extreme events, and its inclusion did not increase the model performance. Furthermore, the results showed that the LR performs similarly to more complex ML algorithms. SVM had the worst performance in all cases. While most of the top-ranked impacting predictors were nearly comparable amongst models, especially CAPE and TCW, we found discrepancies; SAT contributed to RF and NNET but not to LR.