Supplemental Material

Journal of Hydrometeorology

A Causal Inference Model Based on Random Forests to Identify the Effect of Soil Moisture on Precipitation

https://doi.org/10.1175/JHM-D-19-0209.1

© Copyright 2020 American Meteorological Society

Permission to use figures, tables, and brief excerpts from this work in scientific and educational works is hereby granted provided that the source is acknowledged. Any use of material in this work that is determined to be "fair use" under Section 107 of the U.S. Copyright Act or that satisfies the conditions specified in Section 108 of the U.S. Copyright Act (17 USC §108) does not require the AMS’s permission. Publication, systematic reproduction, posting in electronic form, such as on a website or in a searchable database, or other uses of this material, except as exempted by the above statement, requires written permission or a license from the AMS. All AMS journals and monograph publications are registered with the Copyright Clearance Center (http://www.copyright.com). Questions about permission to use materials for which AMS holds the copyright can also be directed to permissions@ametsoc.org. Additional details are provided in the AMS Copyright Policy statement, available on the AMS website (http://www.ametsoc.org/CopyrightInformation).
Supplementary Materials for

A causal-inference model based on Random Forests to identify the effect of soil moisture on precipitation

Lu Li, Wei Shangguan*, Yi Deng, Jiafu Mao, JinJing Pan, Nan Wei, Hua Yuan, Shupeng Zhang, Yonggen Zhang, Yongjiu Dai

Corresponding author: Wei Shangguan (shgwei@mail.sysu.edu.cn)

This PDF file give the reason why we apply Pearson correlation on the whole dataset rather than the train set.
The supplement of the feature selection process.

In the feature selection process, we applied the Pearson correlations on the whole datasets rather than the train sets. We gave the following reasons.

1) Feature selection process is usually based on model (e.g., random forest, logistic regression) and statistical metric (e.g., coefficient). For feature selection based on statistical metric, the correlation is reliable in the statistical sense if the linear correlation is statistically significant. Furthermore, in our perception, no paper used cross-validation for feature selection based on statistical metric. Cross-validation is usually used to avoid overfitting in machine learning methods (used for regression, classification). We indeed used out-of-bag error to avoid overfitting issue in random forest regression.

2) We randomly chose 16 pixels, and calculated both the coefficient on the whole data (see red line in Figure S1) and the coefficient on the training data (70% of the whole data, see blue line in Figure S1). We could figure that the difference between these two coefficient is small, and don’t changed the most important explaining variables selected.
Figure S1. Pearson coefficient of target variable (i.e., POCC) and independent variables (i.e., lagged precipitation and lagged pressure). The red lines is the coefficient on the whole data and the blue line is the coefficient on the training data (70% of the whole data)