26th International Congress on Modelling and Simulation, Adelaide, SA, Australia, 30 Nov–4 Dec 2025
mssanz.org.au/modsim2025

Confronting data leakage in environmental model optimisation

R. Willem Vervoort [a], **Arpit Kapoor** [b], **Rohitash Chandra** [b], **Rajitha Athukorala** [a]

[a]*School of Life and Environmental Sciences, The University of Sydney, NSW 2006*
[b]*School of Mathematics and Statistics, University of NSW, NSW 2052*
Email: willem.vervoort@sydney.edu.au

Abstract: Data leakage, resulting in inflated performance, has been identified as one of the major problems in many published machine learning studies (Kapoor and Narayanan, 2023). Natural resource model problems, are particularly sensitive to data leakage for two major reasons: Most modelling problems are related to open systems with sparse to very sparse sampling of the full data distribution in time and space; and natural systems are complex and signals are often dependent on many cross correlated environmental variables, for example the relationship between rainfall and soil moisture. Furthermore, data related to natural resources can show long term cycles or trends and changes in the relationship between covariates and the predictand over time (Non-stationarity). This all means that regular performance criteria and basic splitting of data between validation and calibration datasets can create biased and over optimistic results (Buchanan, 2024).
As a result, it becomes important to evaluate the typical validation methods used in hydrology and other natural resource sciences. This paper will demonstrate different examples of data leakage in environmental models in space and time predictions and provide recommendations on preventing data leakage in environmental models. In particular, using a simple neural network to forecast a time series (runoff from a catchment) and spatial example (soil moisture, following Filippi et al. (2024)), we demonstrate the impact of different data splitting approaches and use different performance criteria to assess the performance of the models.

Keywords: *Model validation, autocorrelation, hydrology*

References

Buchanan Mark (2024), Don't flock to faulty ai fashion, *Nature Physics* **20**(8), 1220–1220.
URL: *https://doi.org/10.1038/s41567-024-02604-y*

Filippi Patrick, Han Si Yang and Bishop Thomas F. A. (2024), On crop yield modelling, predicting, and forecasting and addressing the common issues in published studies, *Precision Agriculture* **26**(1), 8.
URL: *https://doi.org/10.1007/s11119-024-10212-2*

Kapoor Sayash and Narayanan Arvind (2023), Leakage and the reproducibility crisis in machine-learning-based science, *Patterns* **4**(9), 100804.
URL: *https://www.sciencedirect.com/science/article/pii/S2666389923001599*