#SC5.12

EGU General Assembly 2022

USING R IN HYDROLOGY:

MACHINE LEARNING FOR SPATIO-TEMPORAL MODELLING

Razi Sheikholeslami

1 School of Geography and the Environment, University of Oxford, Oxford, UK

2 Environmental Change Institute, University of Oxford, Oxford, UK

SoGE | School of Geography and the Environment

UNIVERSITY OF OXFORD

What is Machine Learning?

“Machine learning (ML) is the study of computer algorithms that improve automatically through experience. It is seen as a part of artificial intelligence.” Wikipedia

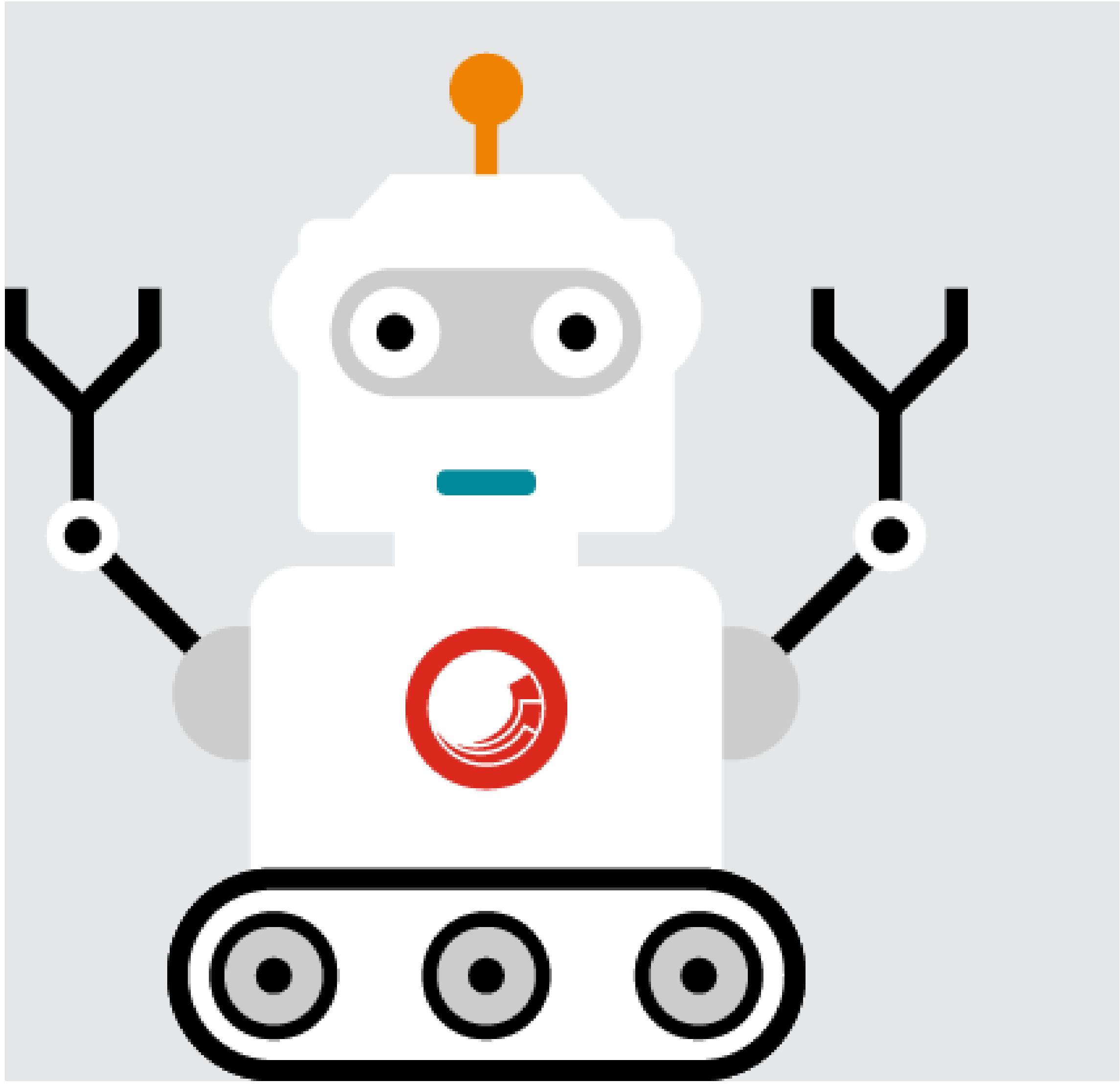
“… a field of science and a powerful technology that allows machines to learn from data and self-improve.” The Royal Society

“Machine-learning algorithms use statistics to find patterns in massive amounts of data. And data, here, encompasses a lot of things—numbers, words, images, … .” MIT Technology Review

The way I see it ..

- Machine learning is a method of data analysis that automates the process of model building.

- It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.



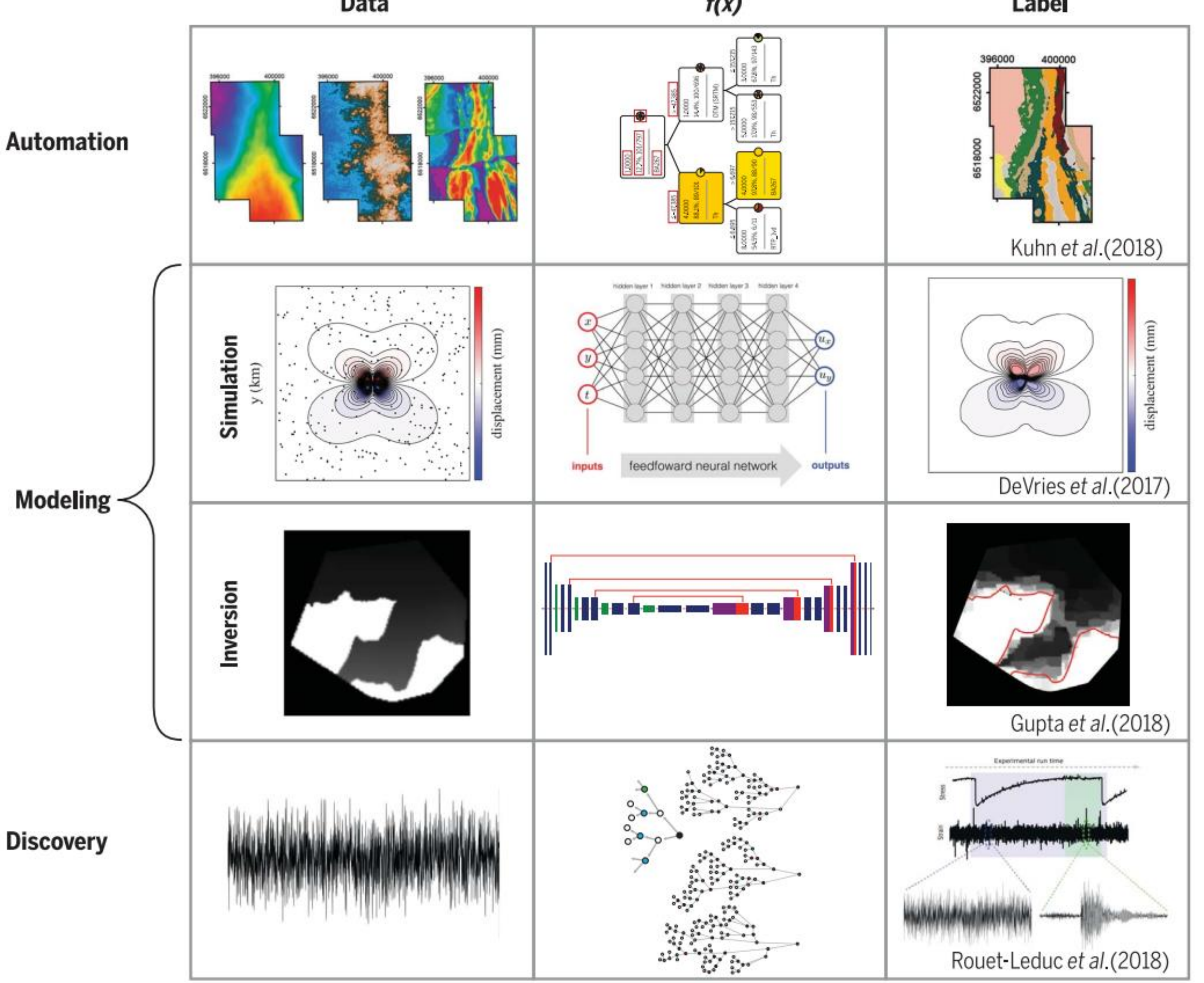
Machine learning approaches are traditionally divided into two broad categories: supervised and unsupervised learning algorithms

- Supervised learning: The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.

- Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).

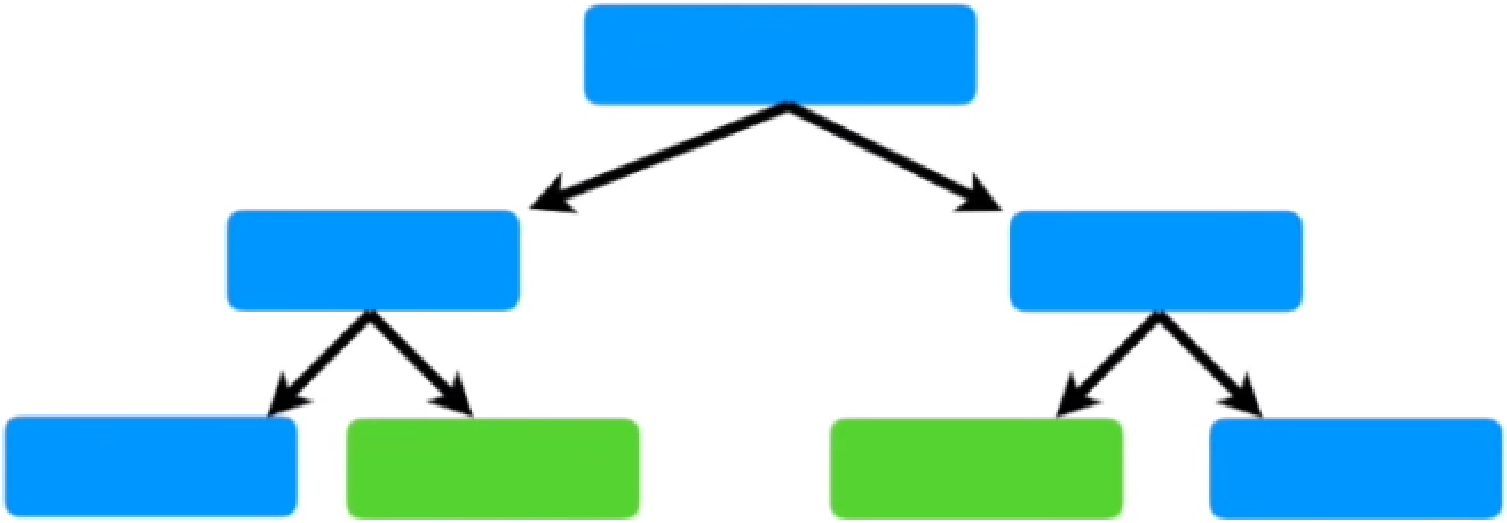
EGU General Assembly 2022

Common Modes of ML



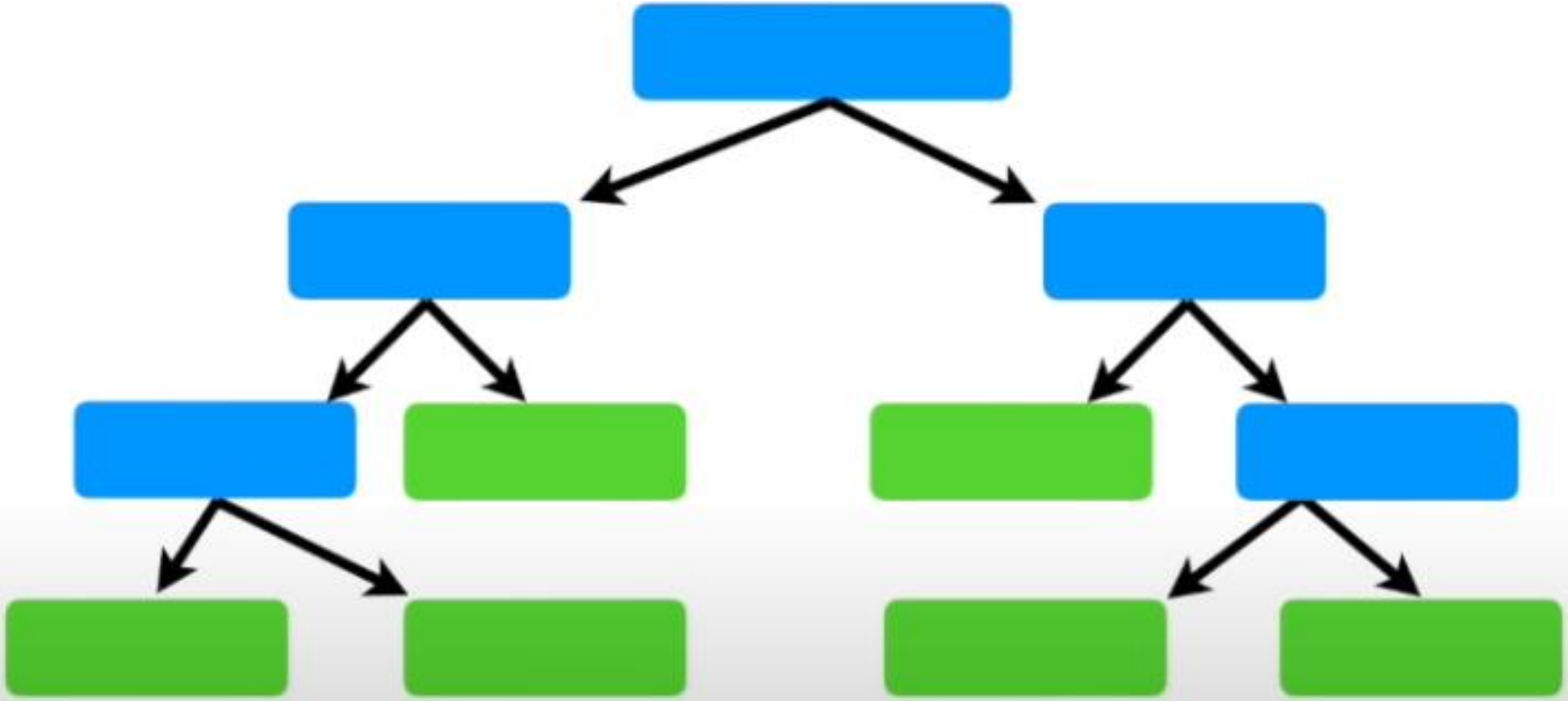
Bergen et al., 2019. Science. https://www.science.org/doi/10.1126/science.aau0323

Random Forests:
A powerful ML Method



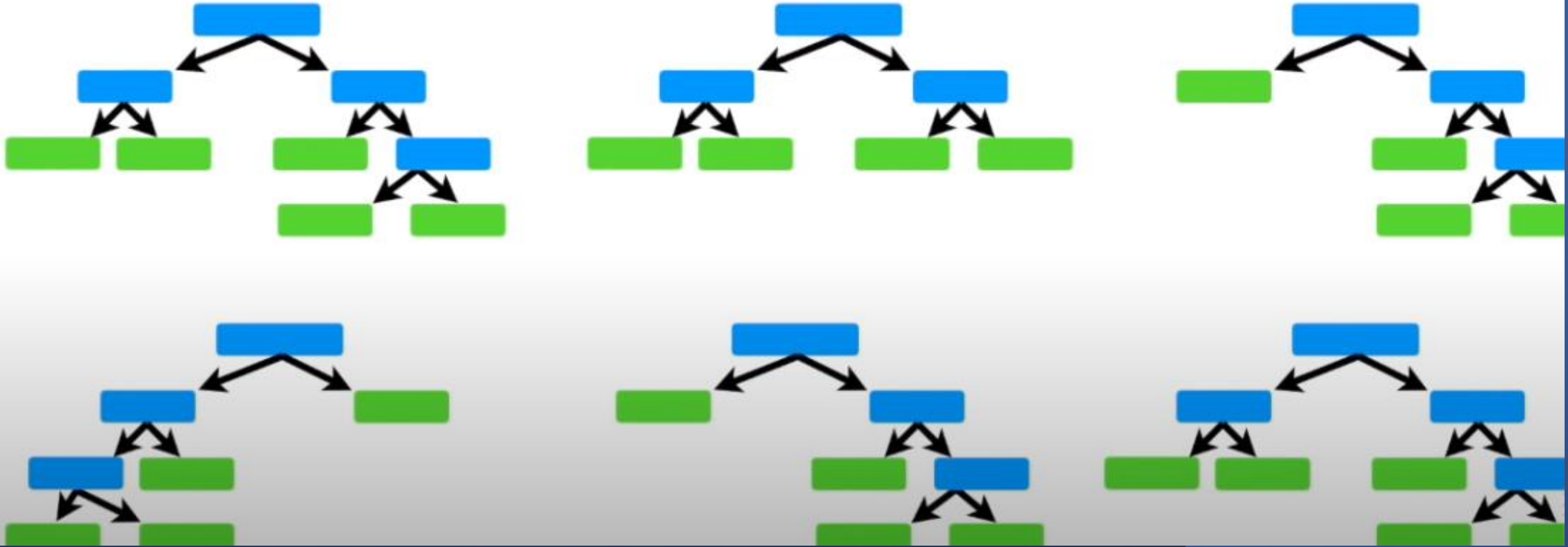
Random forests are made out of decision trees!

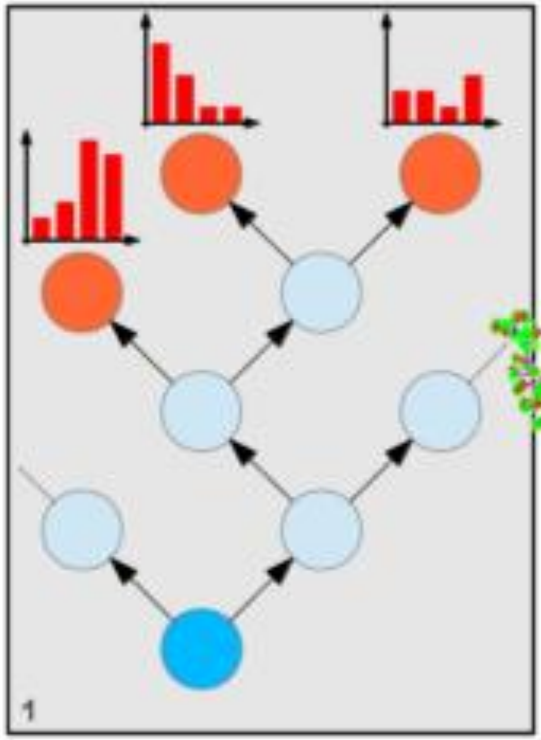
Decision Trees are easy to build, easy to use and easy to interpret...

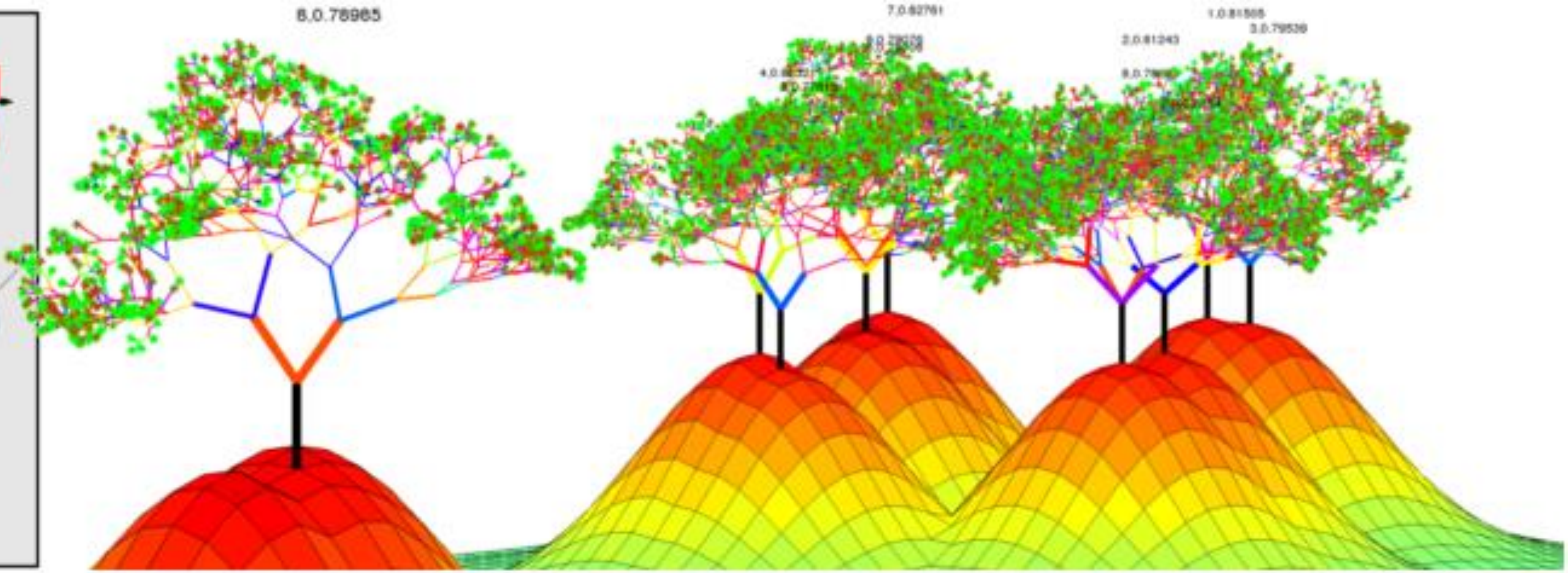


But..!

The good news is that **Random Forests** combine the simplicity of decision trees with flexibility resulting in a vast improvement in accuracy.



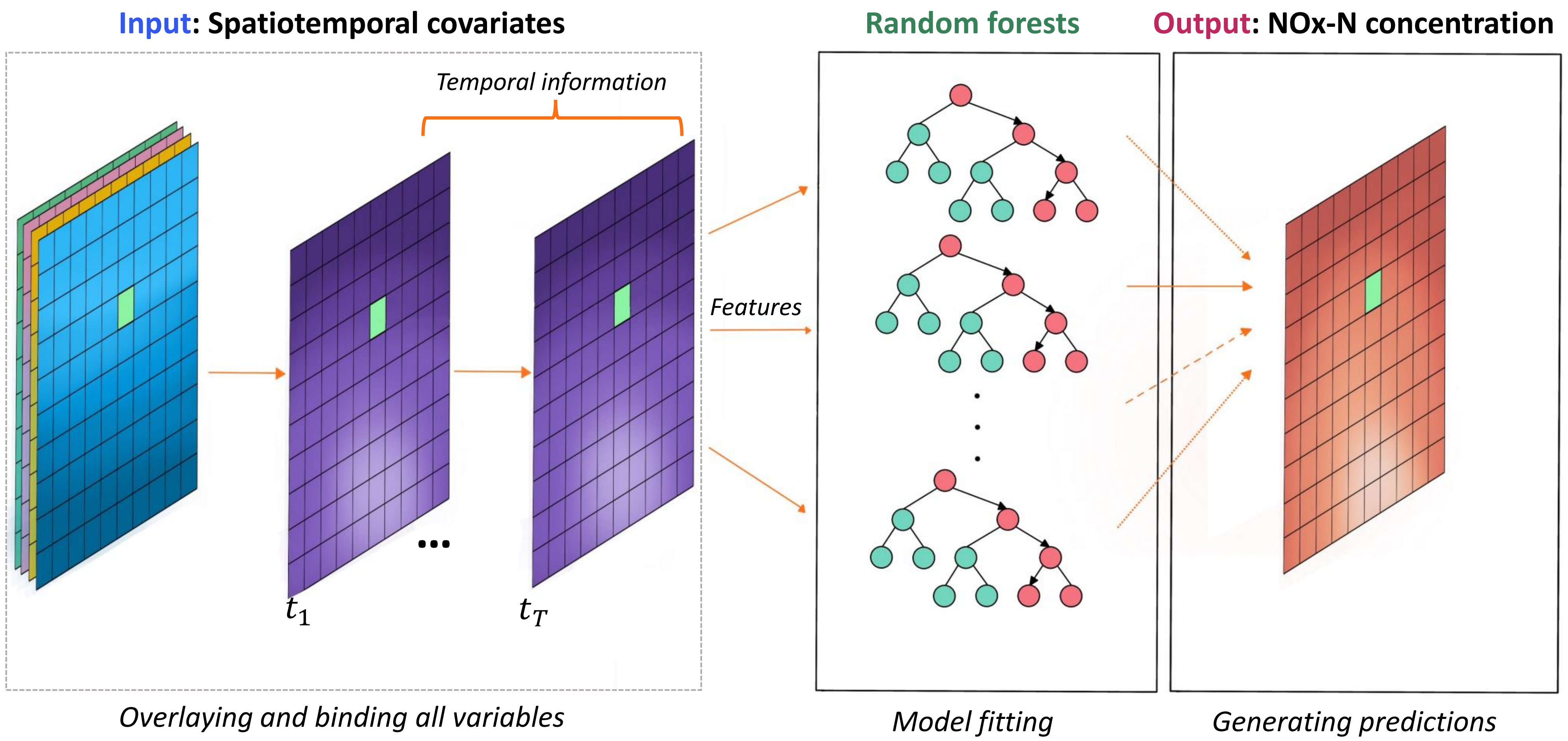




Random Forest Algorithm

★	Fast	(coding vs. runtime vs. interpretation)
★	Replicable	(method well-defined, get same answer)
★	Robust	(insensitive to distributional assumptions, outliers)
★	Predictive ability	(minimal errors, fill in space/time gaps)
★	Covariate effects	(nonlinear, interactions)
--	Uncertainty estimates	(with known properties)

Elements of Spatio-Temporal Modelling with R Using ML



Sheikholeslami and Hall (2021)

Problem Formulation

The assembled dataset is represented by a collection of observations:

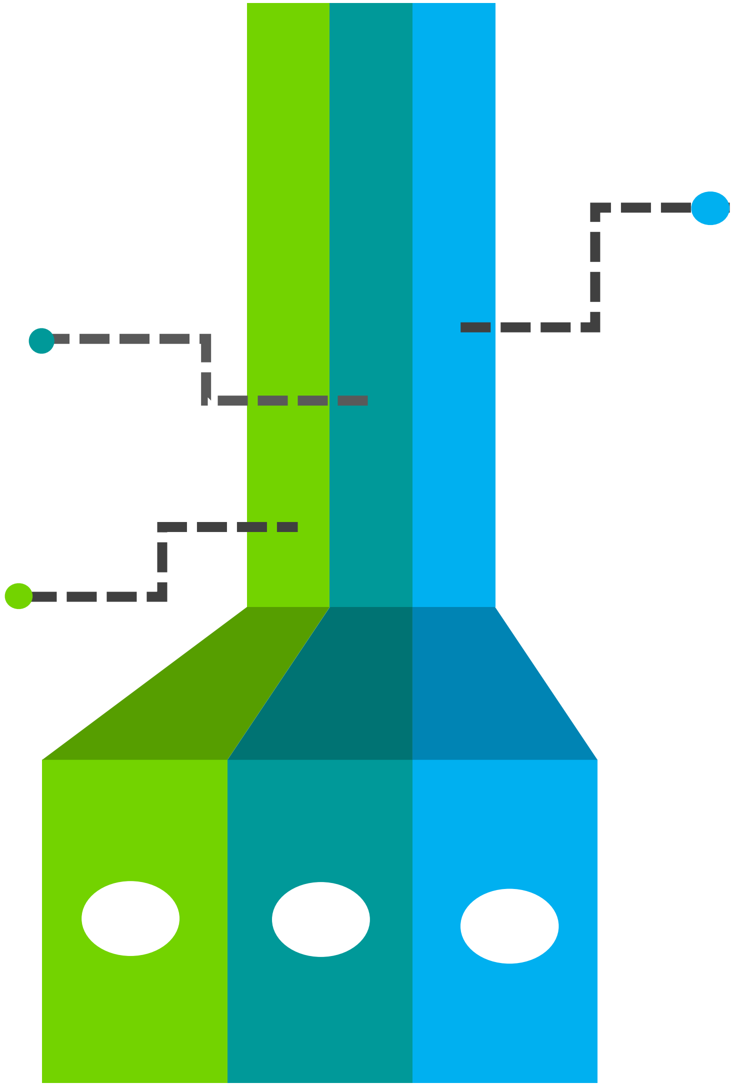
$$\mathbf{Y} = \{y(s_l, t_\tau),\ (s,t) \in \mathcal{S} \times \mathcal{T} \subseteq \mathbb{R}^2 \times \mathbb{R}\}$$

measured at $l = 1,2, \dots, n$ spatial locations and at $\tau = 1,2, \dots, T$ time points over spatial domain $\mathcal{S}$ and temporal domain $\mathcal{T}$.

For example, $y(s,t)$ can be considered as a realization of water quality process, i.e., total nitrogen concentrations.

Representation of Spatio-Temporal Data in R

space-wide, where columns correspond to different spatial features (e.g., locations, regions, grid points, pixels);



long formats, where each record corresponds to a specific time and space coordinate

time-wide, where columns correspond to different time points

Essential R Package

dplyr : data-wrangling spatio-temporal data – in particular filtering, sorting, selecting variables and creating new variables

ggmap : plotting of regional maps

gstat: inverse distance weighting, fitting spatiotemporal semivariograms, and spatiotemporal kriging

spacetime: creating and handling spatio-temporal objects

sp: classes and methods for spatial data;

raster: reading, writing, manipulating, analyzing and modeling of spatial data.

geoR: functions for geostatistical data analysis

rgdal: provides bindings to the 'Geospatial' Data Abstraction Library ('GDAL') and access to projection/transformation operations

ncdf4: reading from, writing to, and creation of netCDF files

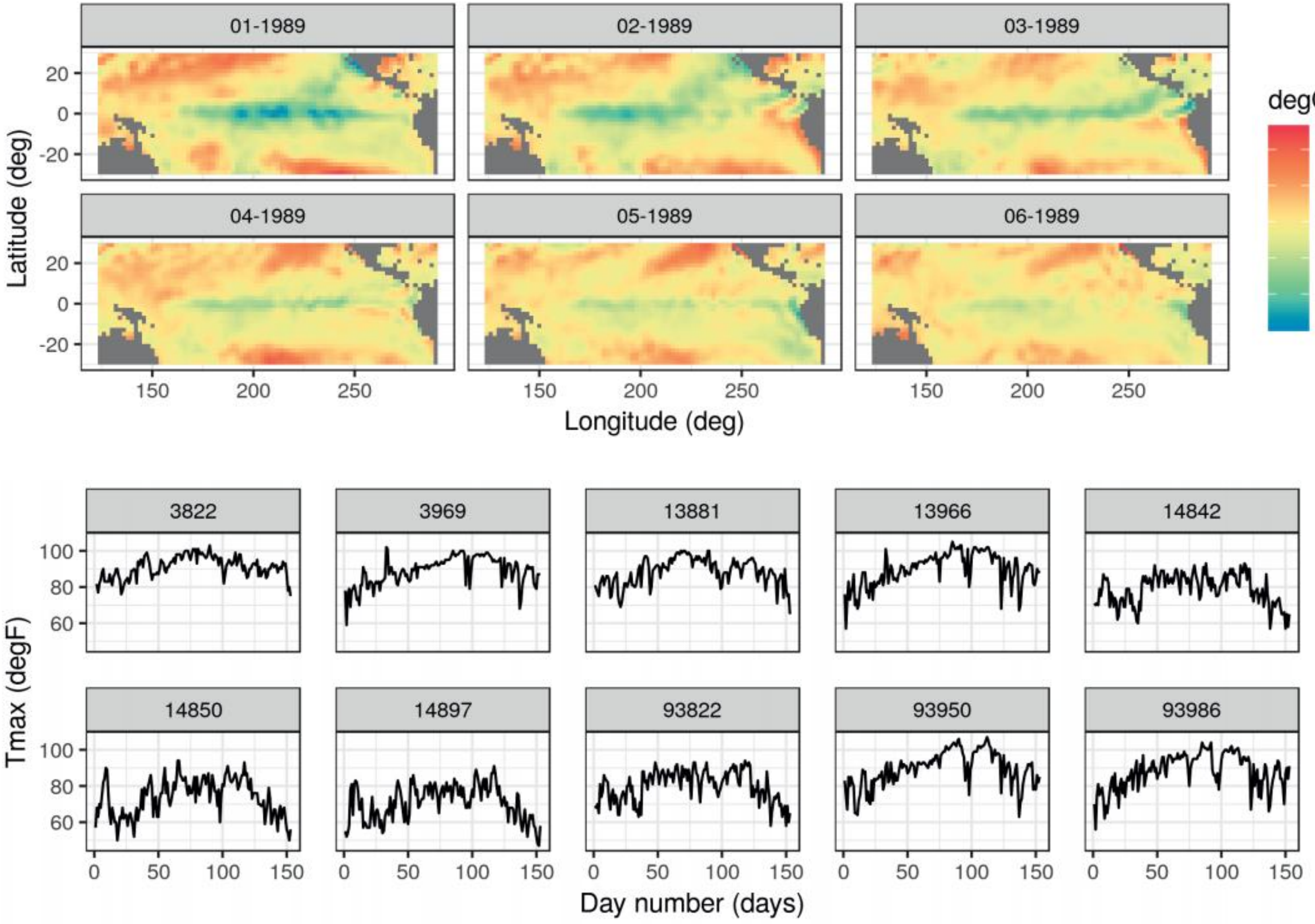
ranger: a fast implementation of Random Forests

Caret: functions for training and plotting classification and regression models

CAST: spatio-temporal model training and prediction using machine learning

Visualization of Spatio-Temporal Data in R

- Spatio-temporal visualization in R generally proceeds using one of two methods: the trellis graph or the grammar of graphics. The command plot invokes the trellis graph when **sp** or **spacetime** objects are supplied as arguments. The commands associated with the package **ggplot2** invoke the grammar of graphics. The data objects frequently need to be converted into a data frame in long format for use with ggplot2, which we often use throughout this book.



Remarks..

- Recent experiments applying machine learning to hydrological modelling indicate that there is significantly more information in large-scale hydrological data sets than hydrologists have been able to translate into theory or models.

- Instead of predicting the quantities of interest directly, machine learning can predict distributional representations (e.g., probabilistic, fuzzy, etc.) directly from input data.

- Towards theory-informed machine learning algorithms

- Where an ML model does outperform relative to a given process-based model, we can conclude that the process-based model does not take advantage of the full information content of the input/output data. At the very least, such cases indicate that there is potential to improve the process-based model(s).

An Example..

THANK YOU

Email:
razi.sheikholeslami@ouce.ox.ac.uk

Twitter: @RaziOptimus

References

Bakar, K. S., & Sahu, S. K. (2015). spTimer: Spatio-temporal Bayesian modeling using R. *Journal of Statistical Software*, *63*, 1-32.

Bergen, K. J., Johnson, P. A., de Hoop, M. V., & Beroza, G. C. (2019). Machine learning for data-driven discovery in solid Earth geoscience. Science, 363(6433), eaau0323.

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, *6*, e5518.

Sheikholeslami, R. and Hall, J.: The role of livestock in nitrogen pollution of large river basins: A machine learning-based assessment, EGU General Assembly 2022, Vienna, Austria, 23–27 May 2022, EGU22-8223, https://doi.org/10.5194/egusphere-egu22-8223, 2022

Sheikholeslami, R., & Hall, J. W. (2021, December). A machine learning approach to identifying the key factors influencing global water quality. In AGU Fall Meeting 2021. AGU.

Wikle, C. K., Zammit-Mangion, A., & Cressie, N. (2019). *Spatio-temporal Statistics with R*. Chapman and Hall/CRC.