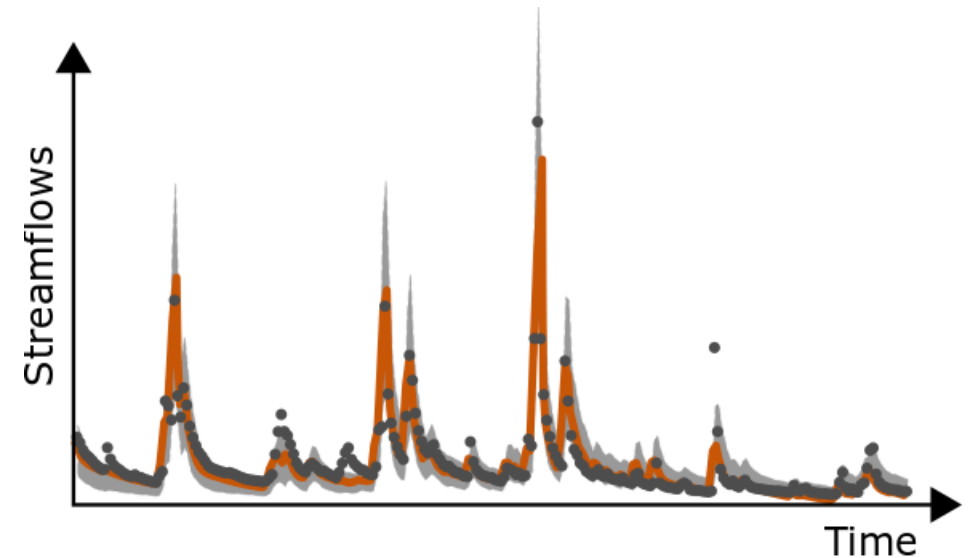


Learning from one's errors: A data-driven approach for mimicking an ensemble of hydrological model residuals

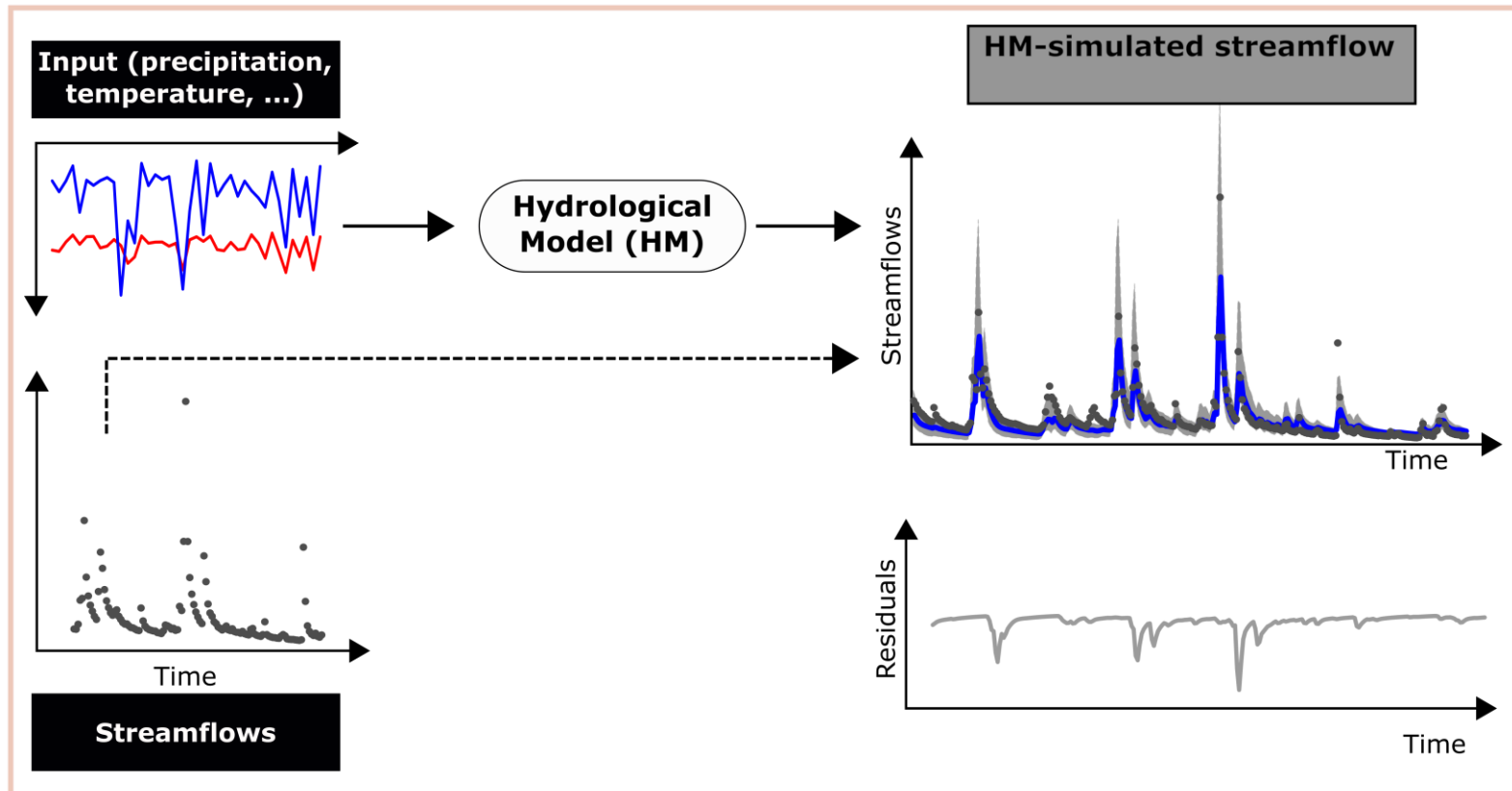
John M. Quilty & Anna E. Sikorska-Senoner

john.quilty@uwaterloo.ca, anna.senoner@geo.uzh.ch



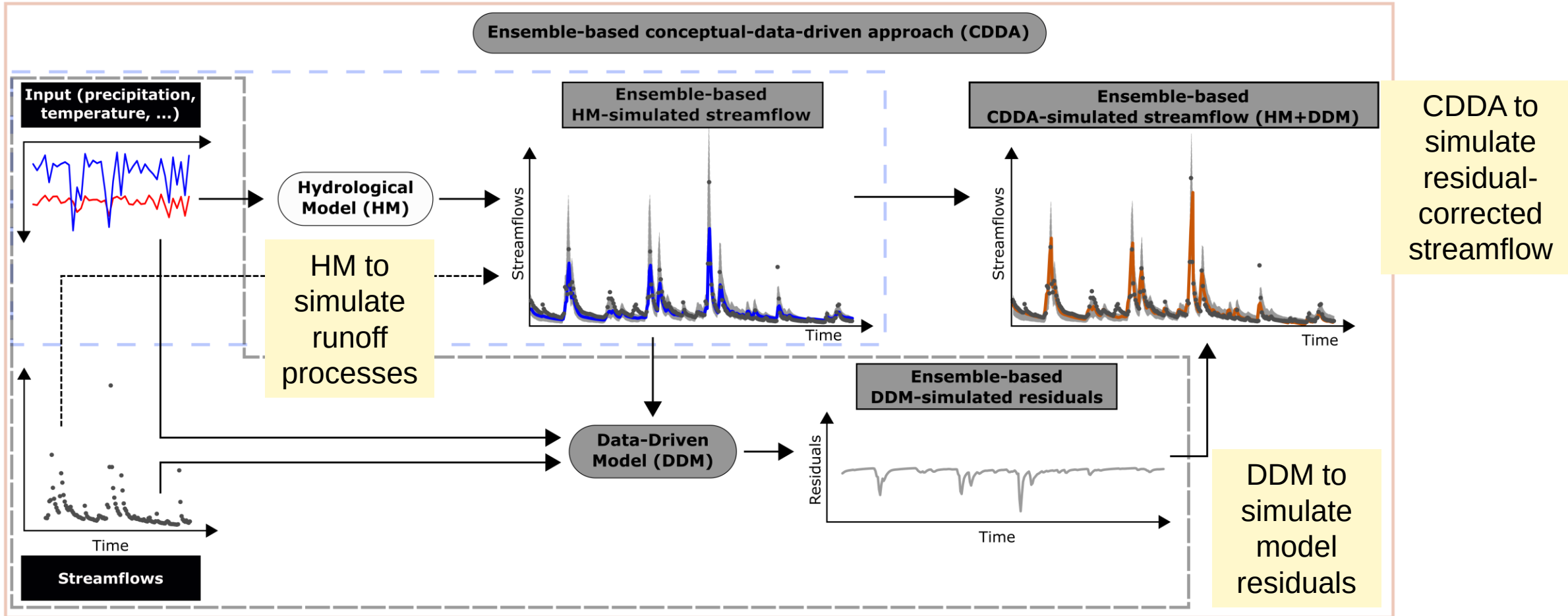
Problem overview

- Hydrological models suffer from errors (residuals) that persist even after the calibration
- Different approaches have been proposed to tackle these errors and some of them require making statements on model error properties (e.g., Gaussian distribution)



Novel concept: coupled ensemble-based streamflow simulation

- A novel concept is introduced: conceptual-data-driven approach (CDDA)
- Hydrological model (HM) coupled with a data-driven model (DDM) into ensemble-based framework



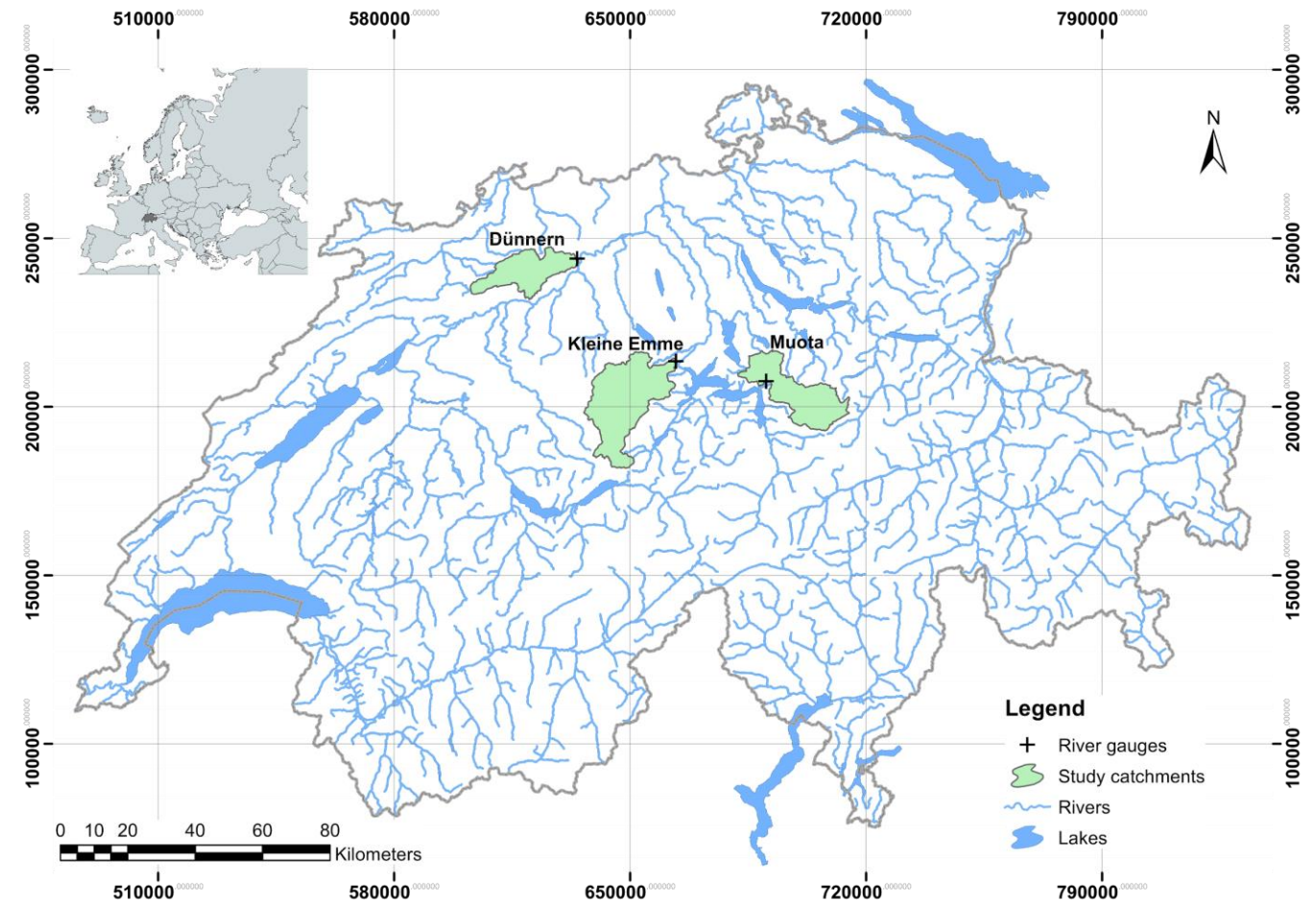
Sikorska-Senoner & Quilty 2021

Experimental set-up

- 3 Swiss catchments
 - Dünnern, Kleine Emme, Muota
- HM = HBV model at a daily time step
- 1000 parameter sets for the HBV
- Record length: 1981-2014

Evaluation based on:

- CRPS (mean continuous ranked probability score)
- MAE (Mean absolute error)



Sikorska-Senoner & Quilty 2021

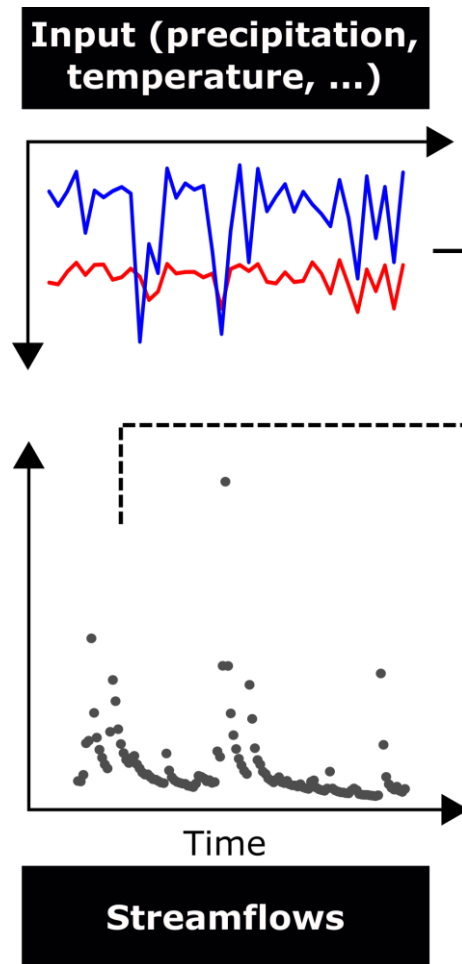
Tested DDMs

- Linear:
 - Multiple Linear Regression (MLR)
- Nonlinear:
 - k Nearest Neighbours Regression (KNN)
 - Second-Order Volterra Series Model (SOV)
 - Artificial Neural Networks (ANN)
 - eXtreme Gradient Boosting (XGB)
 - two variants (all 29 inputs and 6 most important inputs)
 - Random Forests (RF)
 - two variants (all 29 inputs and 6 most important inputs)

Input Variable Selection methods

- Different Input Variable Selection (IVS) methods are coupled with DDMs:
 - Categorized based on linear or nonlinear, model-free or implicit
 - Model-free approaches require IVS external to the model, implicit IVS is performed during model training
 - Linear (model-free):
 - Partial correlation input selection (PCIS) → MLR
 - Nonlinear (model-free):
 - Edgeworth approximations-based conditional mutual information (EA) → KNN, SOV, ANN
 - Nonlinear (implicit):
 - Both XGB and RF contain an implicit IVS mechanism (i.e., external IVS not necessary)

Explored input variables for DDMs

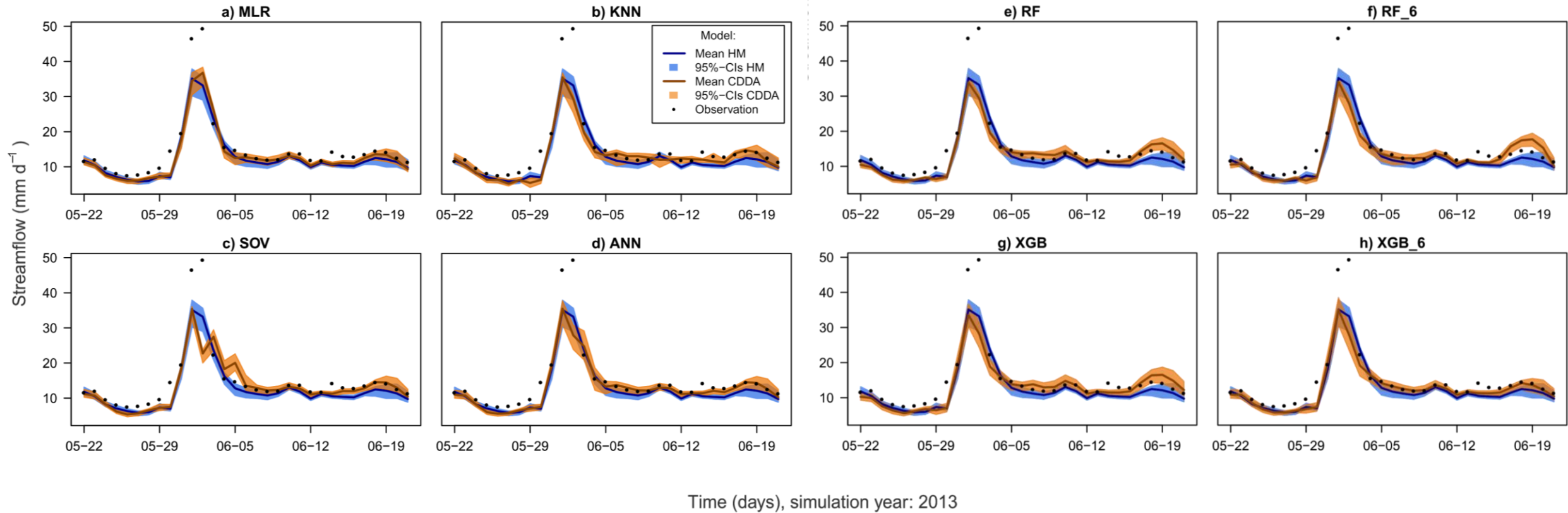


Time-lagged versions of:

- Air temperature ($t-0$ to $t-9$)
- Precipitation ($t-0$ to $t-9$)
- Streamflow ($t-1$ to $t-9$)

Results: Conceptual-data-driven approach (CDDA)

Example for the Muota catchment

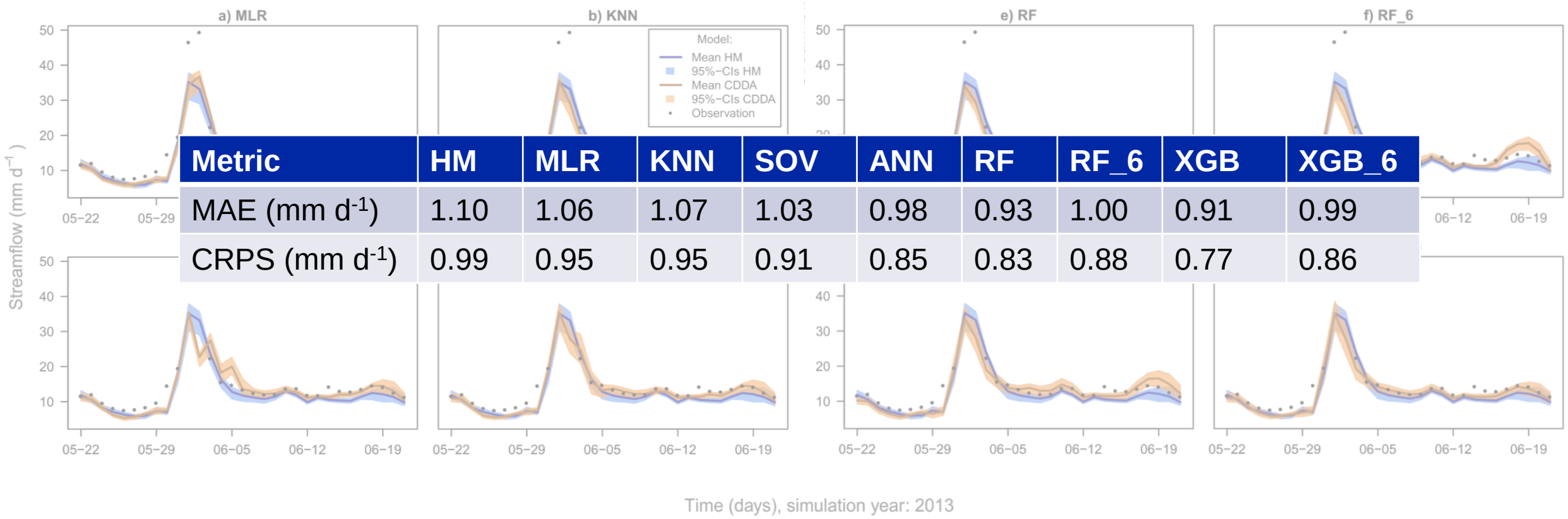


Visual assessment does not give a clear answer on which DDM performs best

Sikorska-Senoner & Quilty 2021

Results: Conceptual-data-driven approach (CDDA)

Example for the Muota catchment



Deterministic and ensemble metrics clearly indicate the benefit of a DDM to simulate residuals with RF and XGB performing the best

Sikorska-Senoner & Quilty 2021

Importance of input variables for DDMs

Most important inputs:

- Precipitation → t-0 to t-3
- Streamflow → t-1 to t-3

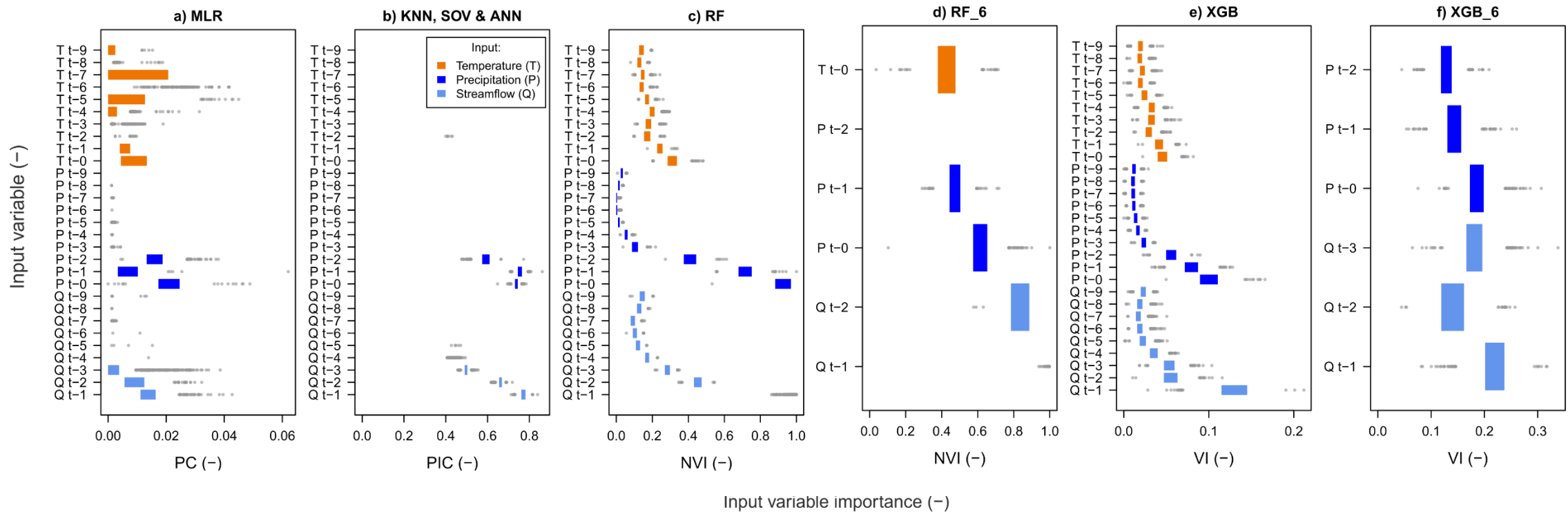
Importance Measure (higher value = more important):

PC: Partial correlation

PIC: Partial informational correlation

NVI: Normalized variable importance

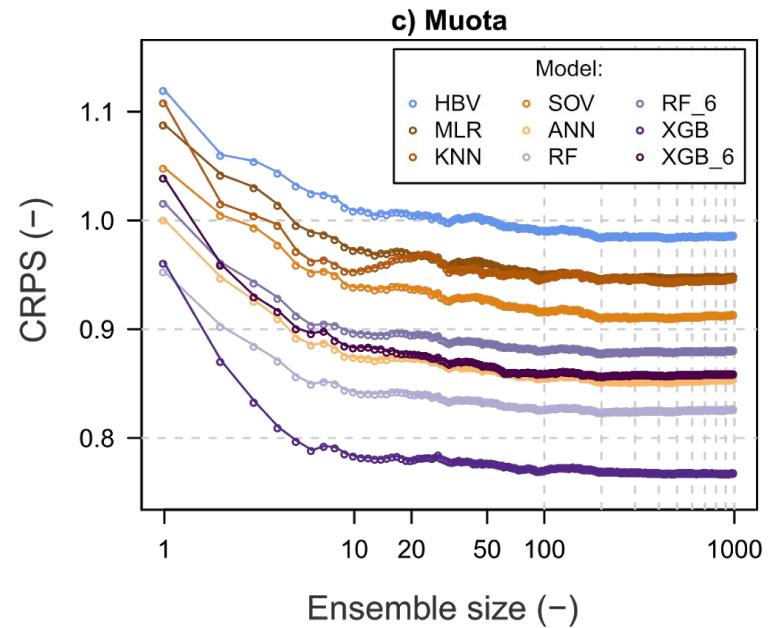
VI: Variable importance



Boxplots indicate input variable importance across ensemble members

Sikorska-Senoner & Quilty 2021

The effect of the ensemble size on CRPS



As few as 100 ensemble members selected out of 1000 provide a similar model performance to the full ensemble with 1000 members

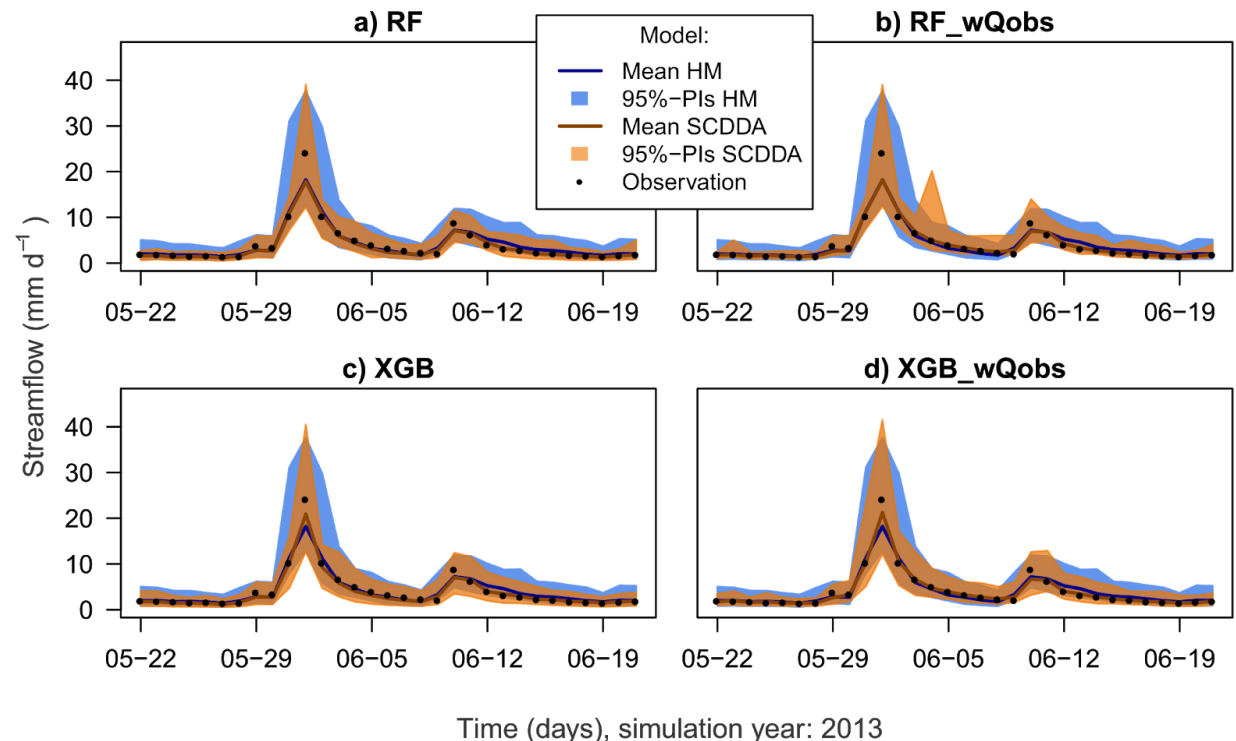
Take home message

- A novel **ensemble-based** conceptual-data-driven approach (CDDA) is proposed;
- CDDA couples a hydrological model (HM) to simulate the runoff process with a data-driven model (DDM) to simulate residuals of the HM;
- Perturbations to HM parameters to generate initial ensemble, DDM coupled with each ensemble member to generate CDDA ensemble
- The CDDA is very flexible – can be coupled with any HM and any DDM.
- Among 8 variants of DDMS tested, **eXtreme Gradient Boosting (XGB)** and **Random Forests (RF)** performed the best
- **Most important inputs** for DDMS were:
 - Precipitation (t-0 to t-3)
 - Streamflow (t-1 to t-3)

Perspectives and work under progress

- Converting the CDDA into a stochastic framework (SCDDA) accounting for additional uncertainty sources (input data, input variable selection, model output) may further improve the performance of ensemble-based simulations
- Exploring SCDDA using different calibration strategies for the HM and DDM (i.e., sequential or simultaneous) may identify trade-offs in terms of performance vs. computational efficiency

SCDDA for the Muota catchment



Literature

Further reading:

Sikorska-Senoner A.E. & Quilty J.M (2021) A novel ensemble-based conceptual-data-driven approach for improved streamflow simulations, Environmental Modelling and Software, in review.

Quilty J.M & Sikorska-Senoner A.E. (2021) A stochastic conceptual-data-driven framework for improved hydrological simulations, Environmental Modelling and Software, in prep.

Contact and further information

John M. Quilty

University of Waterloo, Canada

john.quilty@uwaterloo.ca

<https://uwaterloo.ca/civil-environmental-engineering/profile/jquilty>

Anna E. Sikorska-Senoner

University of Zurich, Switzerland

anna.senoner@geo.uzh.ch

www.annasikorska.eu

