AI-enabled ModEx: A scalable AI workflow to efficiently calibrate PFLOTRAN process models

EMSL UP # 60592

2023 RemPlex Summit Training Workshop

Maruti Mudunuru
Earth Scientist, PNNL

Date: Nov-17-2023

PNNL-SA-184045

Background and relevance

- **Model –** The “PFLOTRAN reaction sandbox” provides users with a venue for implementing user-defined reactions with minimal code
- **Data –** Relevance -- Reaction networks models developed within the reaction sandbox can be leveraged to better understand carbon cycling
- **AI-enabled ModEx –** The PFLOTRAN reaction sandbox models and experimental/field data integration can be performed using AI/ML

PFLOTRAN / code / pflotran

reaction_sandbox_flexbiohill.F90

Pull requests | Check out

PFLOTRAN is an open source, state-of-the-art massively parallel subsurface flow and reactive transport code.

Source | master | 7642e2e | Full commit

pflotran / src / pflotran / **reaction_sandbox_flexbiohill.F90** | Edit

```
module Reaction_Sand_FlexBioHill_class

#include "petsc/finclude/petscsys.h"
  use petscsys

  use Reaction_Sandbox_BioHill_class
  use PFLOTRAN_Constants_module

  implicit none

  private

  type, public, &
    extends(reaction_sandbox_biohill_type) :: reaction_sandbox_flexbiohill_type
    PetscReal :: k_max
    PetscReal :: K_Aaq_n
    PetscReal :: K_Baq
    PetscReal :: I_Caq
    PetscReal :: yield
    PetscReal :: k_decay
    PetscReal :: n
    PetscBool :: molarity_units
    PetscReal, pointer :: stoich(:)
  contains
    procedure, public :: ReadInput => FlexBioHillReadInput
    procedure, public :: Setup => FlexBioHillSetup
    procedure, public :: Evaluate => FlexBioHillEvaluate
    procedure, public :: Destroy => FlexBioHillDestroy
  end type reaction_sandbox_flexbiohill_type

  public :: FlexBioHillCreate

contains

! ************************************************************************** !

function FlexBioHillCreate()
  !
  ! Allocates flexible biodegradation reaction object.
  !
  implicit none

  class(reaction_sandbox_flexbiohill_type), pointer :: FlexBioHillCreate
```

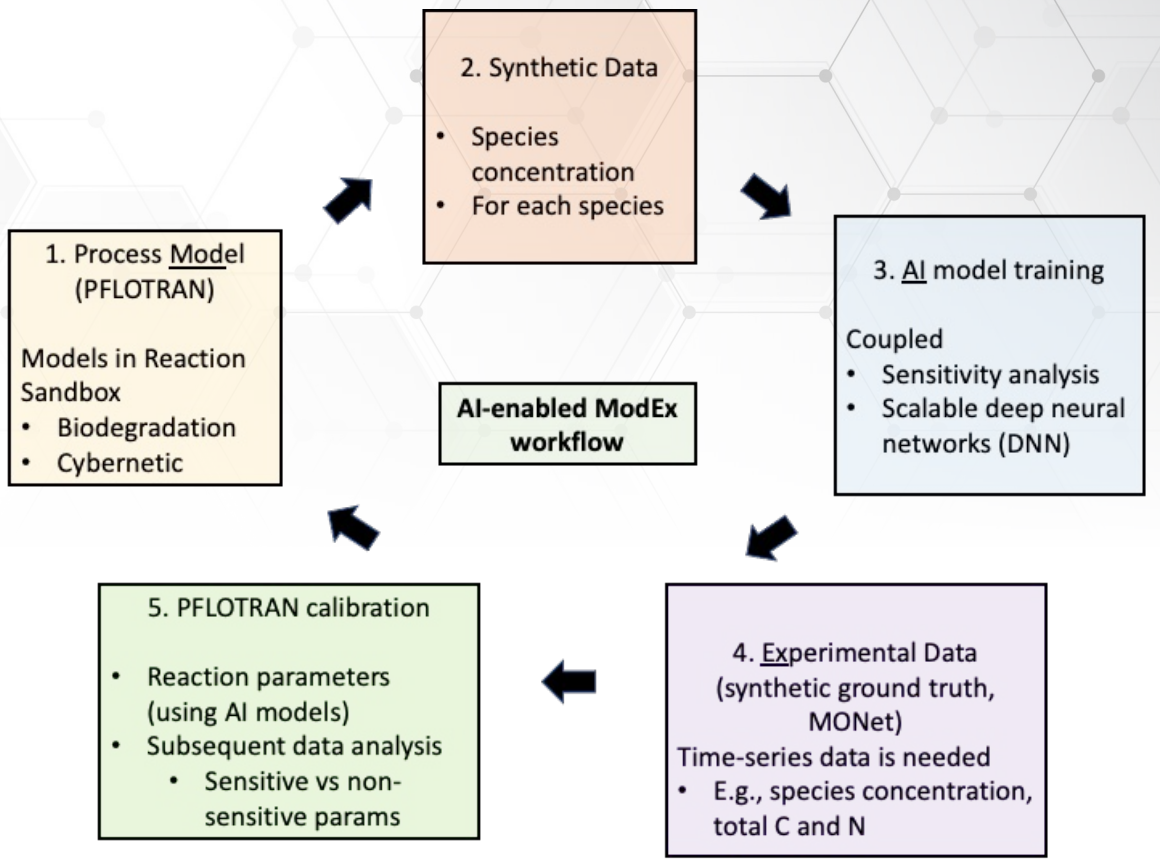
Example – Reaction Sandbox F90 code for implementing user-defined reactions

Example – PFLOTRAN input deck for numerical simulations

```
# Description: Reaction Sandbox Simple (A + B --> C + D)
# Number of reaction network parameters: D = 7
# Sobol-based GSA: N * (2D+2) samples for first and second-order indices

#=========================================================================
SIMULATION
  SIMULATION_TYPE SUBSURFACE
  PROCESS_MODELS
    SUBSURFACE_TRANSPORT transport
      MODE GIRT
    /
  /
END
#=========================================================================

SUBSURFACE

#========================== constraints ==================================
# modify these initial concentration
CONSTRAINT initial
  CONCENTRATIONS # [mol/L]
    Aaq  1.d-3   F
    Baq  5.d-4   F
    Caq  1.d-10  F
    Daq  1.d-10  F
    Eaq  1.d-10  F
  /
  IMMOBILE     # [mol/m^3 bulk]
    Xim  1.d-4
    Yim  1.d-10
  /
END

#========================== chemistry ============================================
CHEMISTRY
  PRIMARY_SPECIES
    Aaq
    Baq
    Eaq
    Caq
    Daq
  /
  IMMOBILE_SPECIES
    Yim
    Xim
  /
```

Why AI-enabled Model-Experiment-Data integration (ModEx)?

- Background
  - Provide calibrated PFLOTRAN reaction process models that can work with field or experimental data (e.g., carbon degradation)
  - Sensitivity analysis to identify important parameters
  - Model-Data integration using AI/ML
- End-to-end workflow outcomes
  - An open-source and user-friendly AI/ML workflow available for users to analyze field/experimental data
  - A skeleton that can be easily adapted for future workflows
    - Other process models relevant to coupled flow, transport, and ERT field data
    - AI/ML modeling for PFLOTRAN users
- Technical readiness plan (proof of concept, development, deployment)
  - Proof-of-concept (TRL-3)
  - AI/ML workflows for PFLOTRAN users (TRL-3)



PFLOTRAN reaction sandbox – Biomass degradation process model

- **Carbon degradation –** Microbially mediated carbon biodegradation reaction with biomass growth and decay over time

$$A_{aq} + 0.25 B_{aq} \rightarrow 0.33 C_{aq} + D_{aq},$$

Electron donor (e.g., carbon compounds)

Electron acceptor (e.g., oxygen)

- **Model –** The rate of biomass growth and decay over time can be modeled as

$$\frac{dX}{dt} = yield_{X_{im}} I_r - k_{decay} X$$

Change in concentration

Yield

Biomass decay

- **Reaction rate –** Monod expression + Hill function

$$I_r = k_{max} X_{im} \frac{A_{aq}^n}{K_{A_{aq}}^n + A_{aq}^n} \times \frac{B_{aq}}{K_{B_{aq}} + B_{aq}} \times \frac{I_{C_{aq}}}{I_{C_{aq}} + C_{aq}}$$

- Process model parameters in this reaction network that are varied
  1. Max Specific Utilization Rate – k_{max}
     - [mole mole^{-1} $_{biomass}$ s^{-1}]
  2. Aaq Half Saturation Constant – K_{Aaq} [M]
  3. Baq Half Saturation Constant – K_{Baq} [M]
  4. Caq Monod Inhibition Constant – K_{Caq} [M]
  5. Yield – $yield_{Xim}$ [mole$_{biomass}$ mole^{-1}]
  6. Biomass Decay Rate Constant – k_{decay} [s^{-1}]
  7. Hill Exponent – n [-]

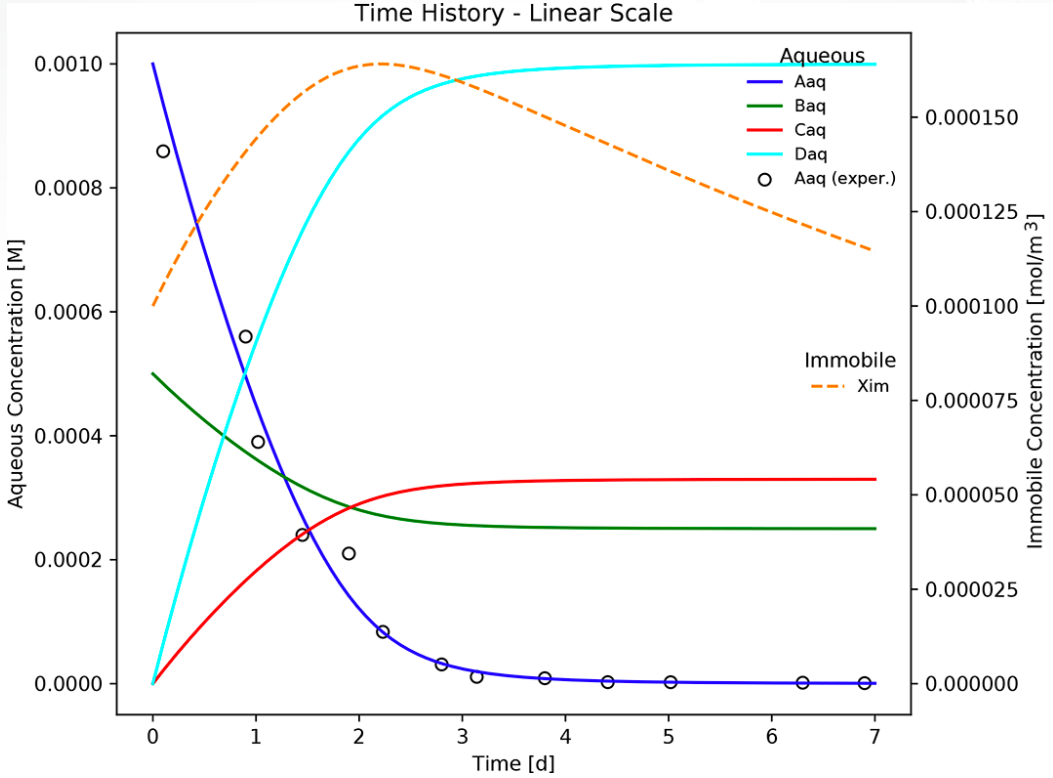
- Inhibitor concentration – I_{Caq} [M] is not varied
- Immobile biomass species – X_{im} [mole$_{biomass}$ m^{-3}$_{bulk}$] is not varied

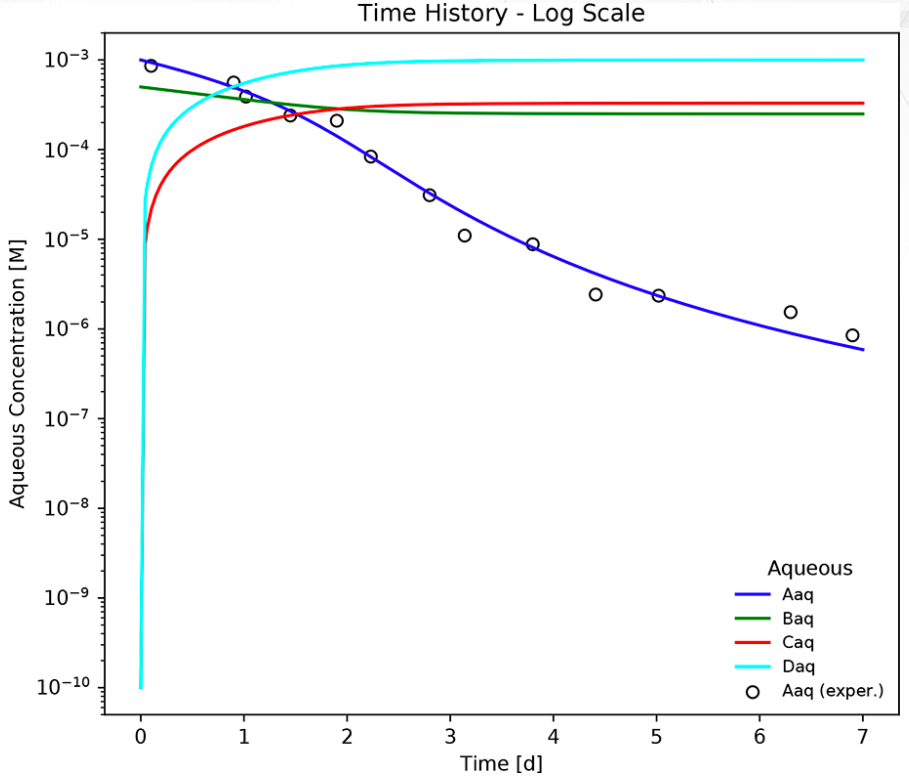
Hammond et al., PFLOTRAN reaction sandbox, 2022

Data requirements – What is needed to calibrate PFLOTRAN reaction sandbox? (1/2)

- **Data** – What are the data requirements for calibrating such reaction network parameters?
  - We need time-series data on species concentration (i.e., organic carbon and major cation/anion concentrations)
  - We will use synthetic time-series to develop and test the workflow

For example, incase of the biomass degradation process model





Data requirements – What is needed to calibrate PFLOTRAN reaction sandbox? (2/2)

- **Data** – What are the data requirements for calibrating such reaction network parameters?
  - We need time-series data on species concentration (i.e., organic carbon and major cation/anion concentrations)
  - We will use synthetic time-series to develop and test the workflow

Table 2. Microbially mediated reaction parameters for the batch biodegradation experiment. n only applies to the reaction incorporating the Hill function (i.e., Eq. 15). M signifies molarity or mole per liter of water.

Parameter	Value	Units
k_{max}	9×10^{-2}	mole $\text{mole}_{biomass}^{-1}$ s^{-1}
$K_{A_{aq}}$	2×10^{-4}	M
$K_{B_{aq}}$	1.25×10^{-5}	M
$I_{C_{aq}}$	2.5×10^{-4}	M
$yield_{X_{im}}$	1×10^{-4}	mole biomass $mole^{-1}$
k_{decay}	$1 \times 1-^{-6}$	$1\ s^{-1}$
n	1.2	–

Table 3. Initial concentrations for the batch biodegradation experiment.

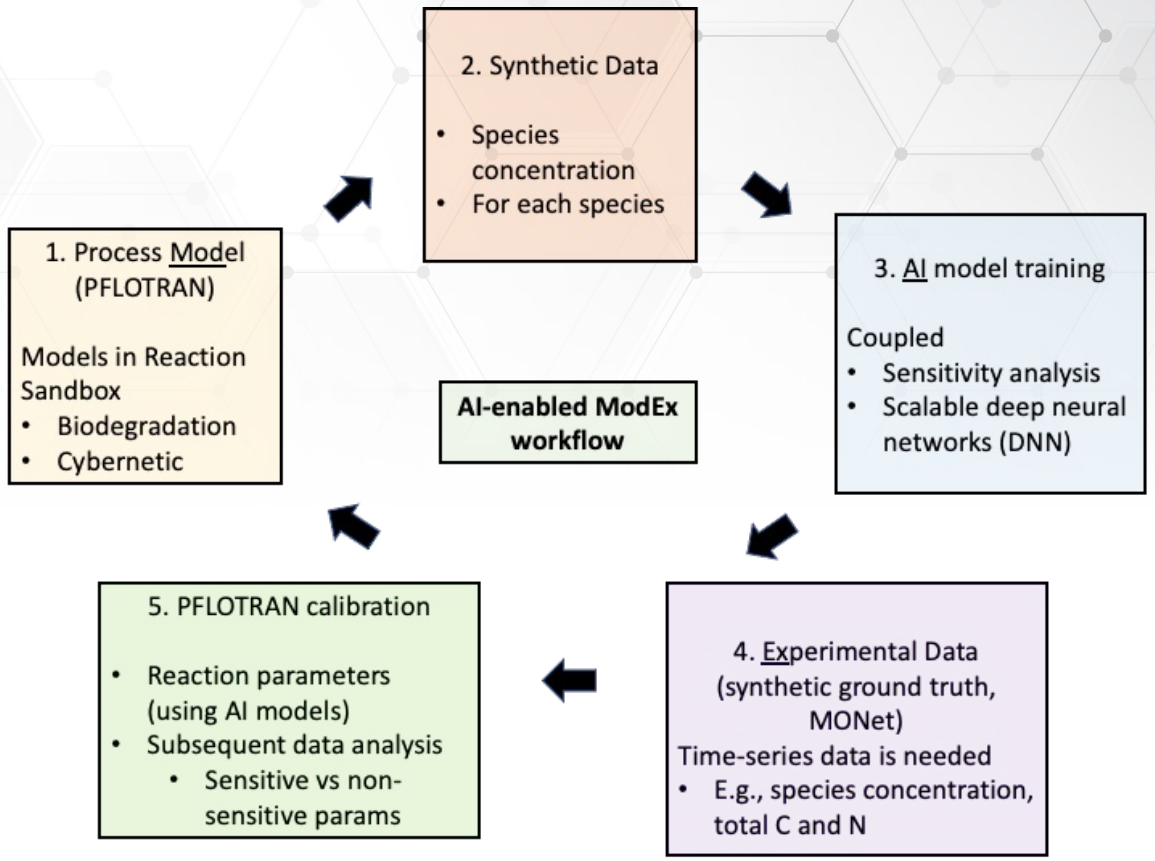
Species	Concentration	Units
A_{aq}	1×10^{-3}	M
B_{aq}	5×10^{-4}	M
C_{aq}	1×10^{-10}	M
D_{aq}	1×10^{-10}	M
X_{im}	1×10^{-4}	mole m_{bulk}^{-3}

Microbially mediated biodegradation reaction with biomass growth and decay over time: Hammond et al., PFLOTRAN reaction sandbox, 2022

AI-enabled ModEx: Overall summary

AI-enabled ModEx pipeline that will be available to PFLOTRAN users has five steps

1. Microbial reaction specifics with synthetic data
   - Examples from PFLOTRAN reaction sandbox
2. PFLOTRAN simulation data generation
   - Sobol-based sampling for realizations
     - N * (2D + 2) realizations = 16000
3. Sensitivity analysis (Local, Global, Obs-based)
   - F-test, Mutual Information
   - Random Forests, SHAPley values
4. Reaction network parameter estimation
   - AI model training
   - With and without noise in data
5. Comparison with synthetic data
   - Performance metrics (e.g., R^2-score, MSE)
   - Predictive uncertainty

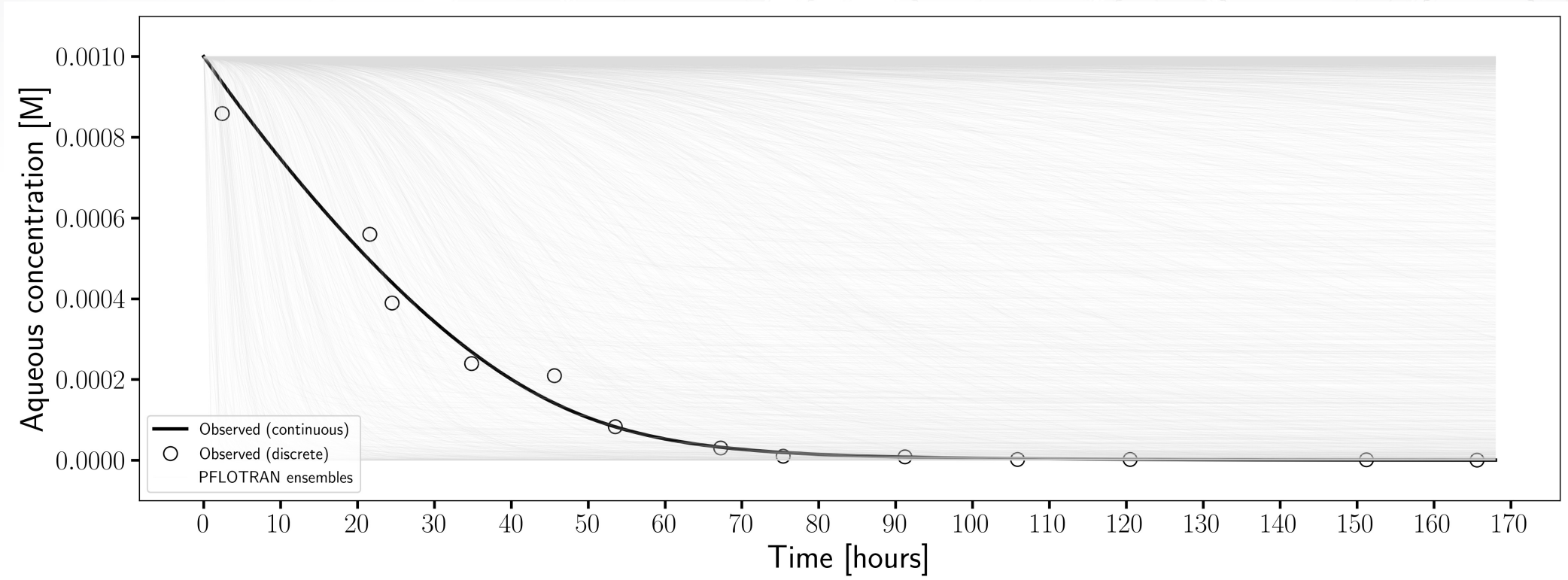


Step-1 and Step-2 – PFLOTRAN simulations

- **Carbon degradation –** Microbially mediated carbon biodegradation reaction with biomass growth and decay over time
- Simulation data generation
  - 16000 Sobol-sampling realizations
    - 15539 realizations ran to completion

Process model parameters in this reaction network

1. Max Specific Utilization Rate
2. Aaq Half Saturation Constant
3. Baq Half Saturation Constant
4. Caq Monod Inhibition Constant
5. Yield
6. Biomass Decay Rate Constant
7. Hill Exponent

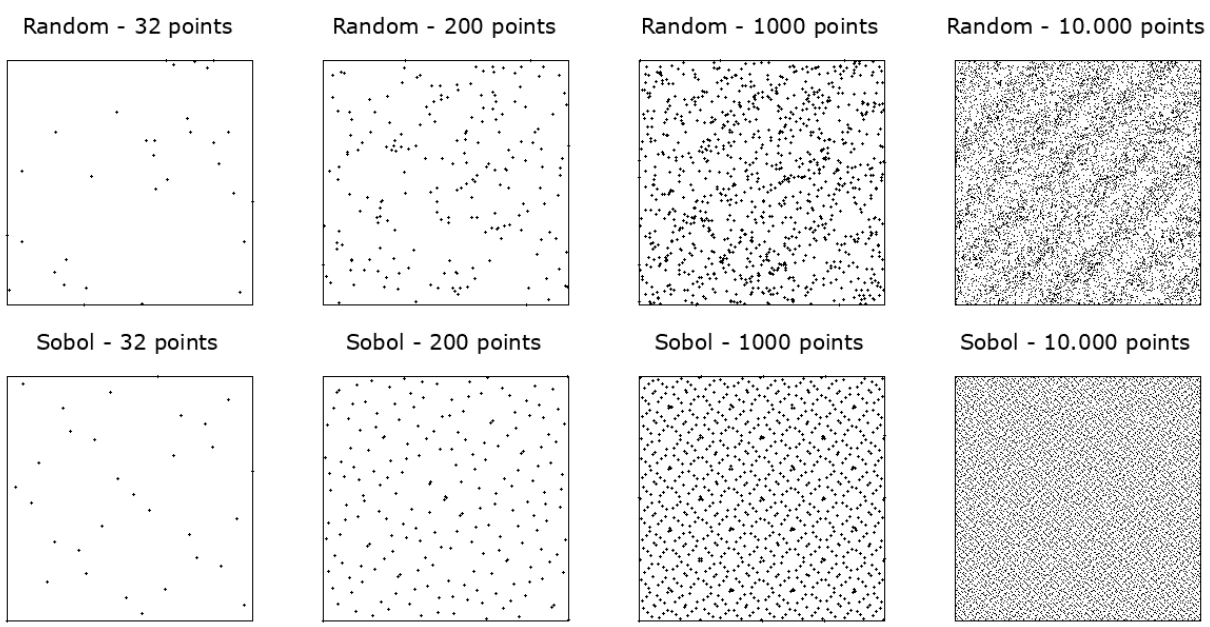


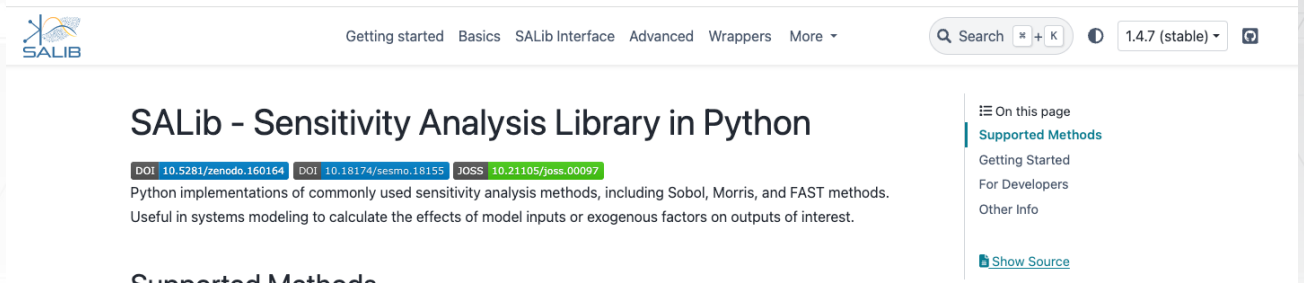
What is a Sobol Sequence?

Sobol sequence is a quasi-random numbers generated by Sobol's algorithm in place of (pseudo-)random numbers.

Quasi-random numbers offer a lower discrepancy (they fill the space of possibilities more evenly) resulting in a faster convergence and more stable estimates of quantities of interest and sensitivities

SALib: https://salib.readthedocs.io/en/latest/

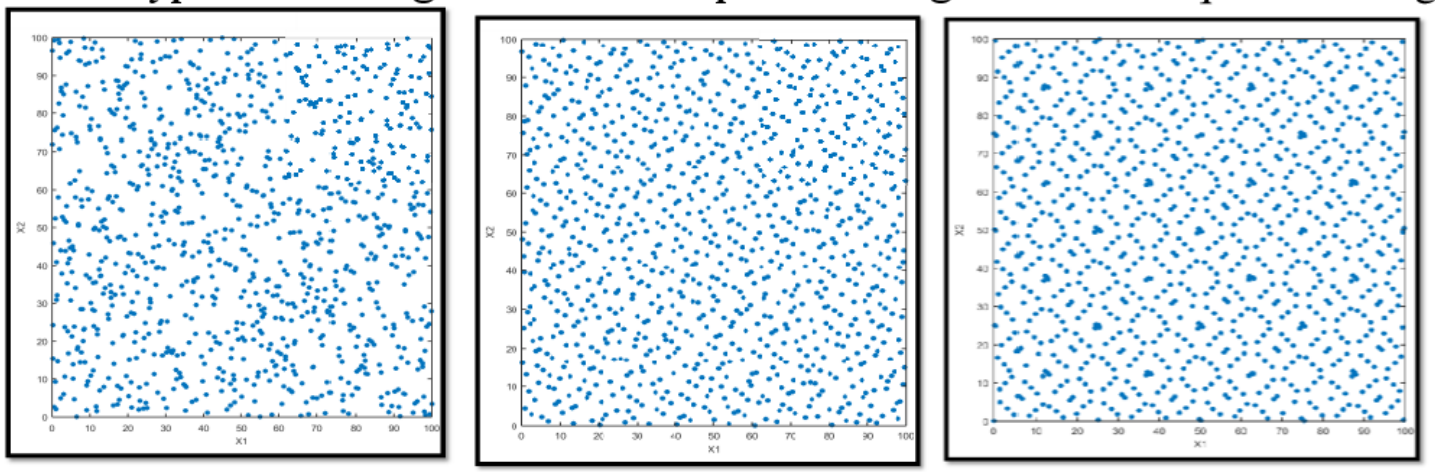




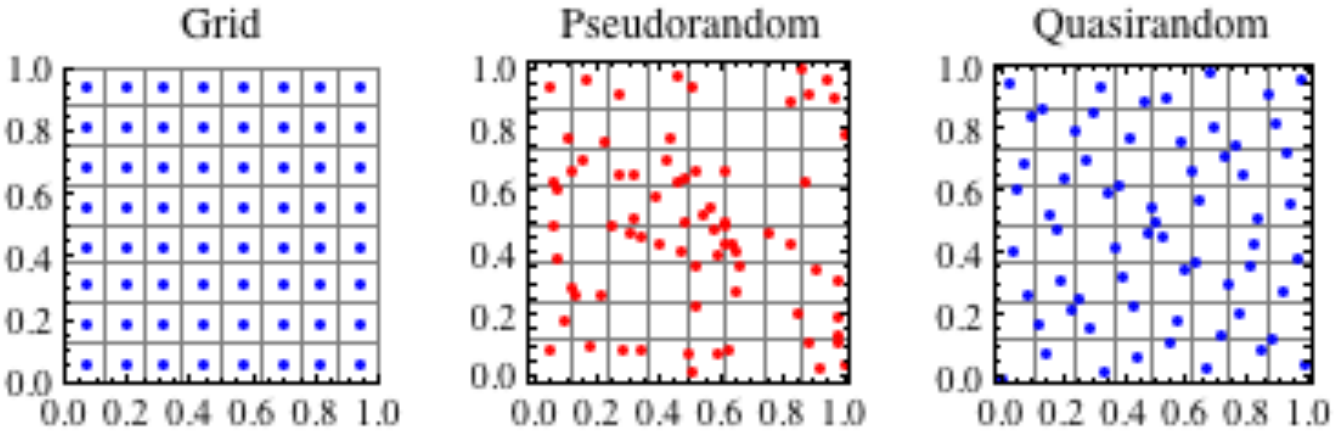
Alternate sampling strategies and low-discrepancy sequences

Low-discrepancy sequences are an excellent tool for evenly sampling a sample space. They generalize well to any real space, if you use an appropriate mapping.

Θ Latin hypercube design Θ Halton sequence design ΘSobol sequence design



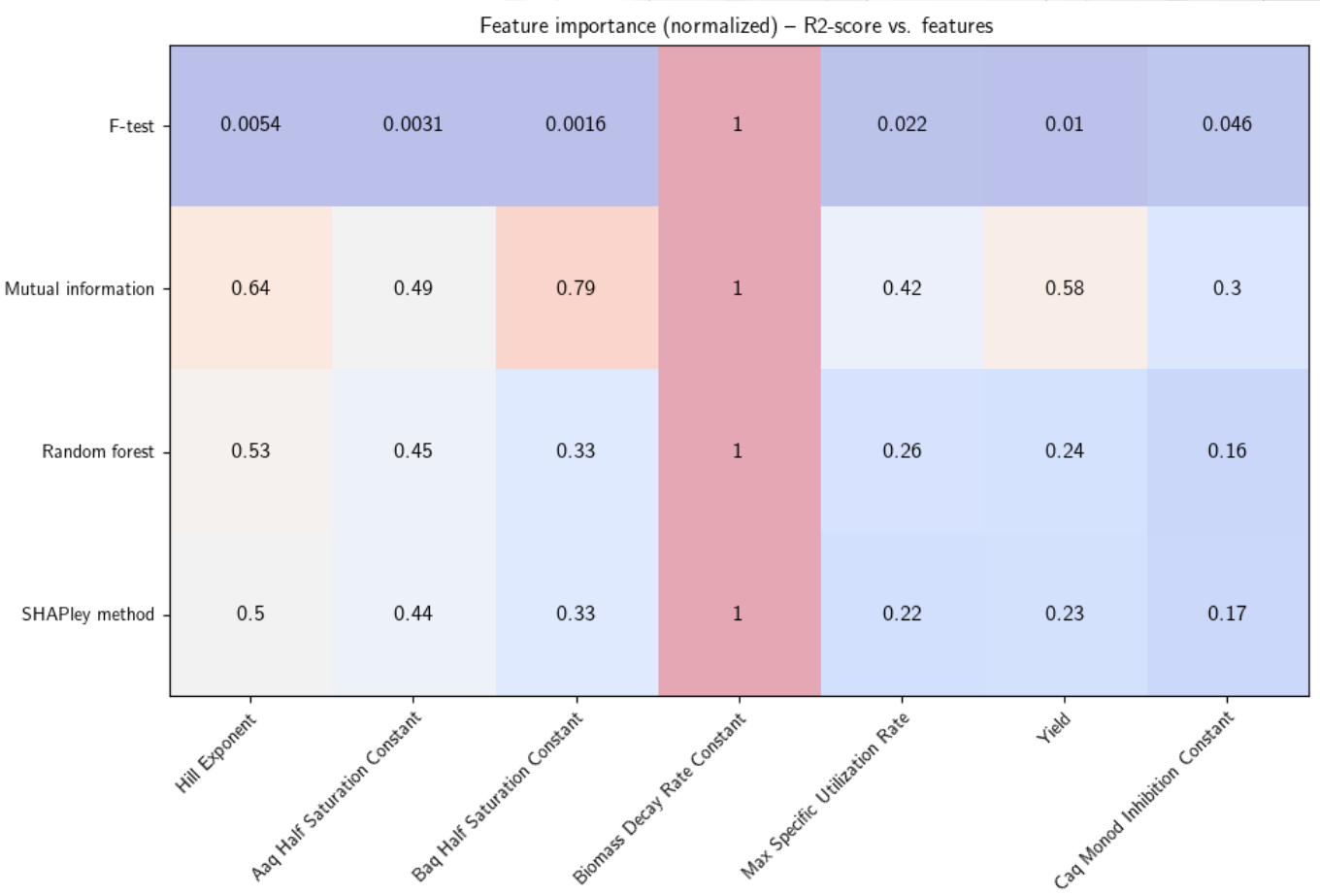
Alternative 2D sequences



Step-3: Comprehensive sensitivity analysis (1/3)

- **Local sensitivity analysis**
  - F-test
  - Mutual information (MI)
- **Global sensitivity analysis**
  - Averaged local sensitivities
- **Sensitivity analysis representative of observations**
  - F-test
  - MI
  - Random Forest
  - SHAPley values
  - Performance metrics
    - R^2, NSE, logNSE, KGE, mKGE, npKGE

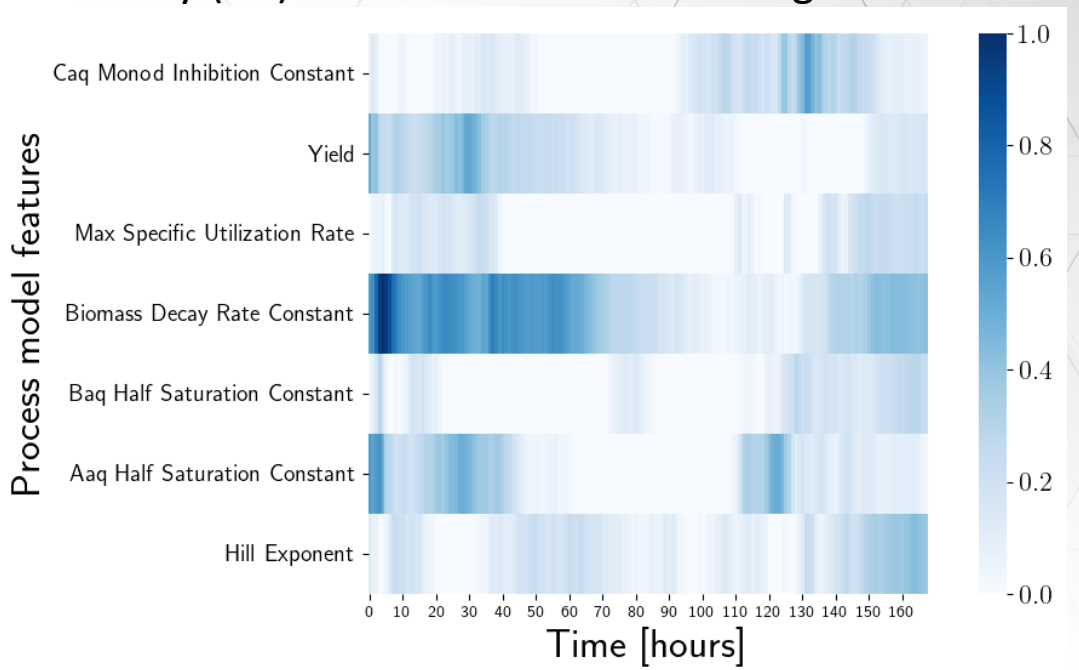
Summary of sensitivity analysis using various methods

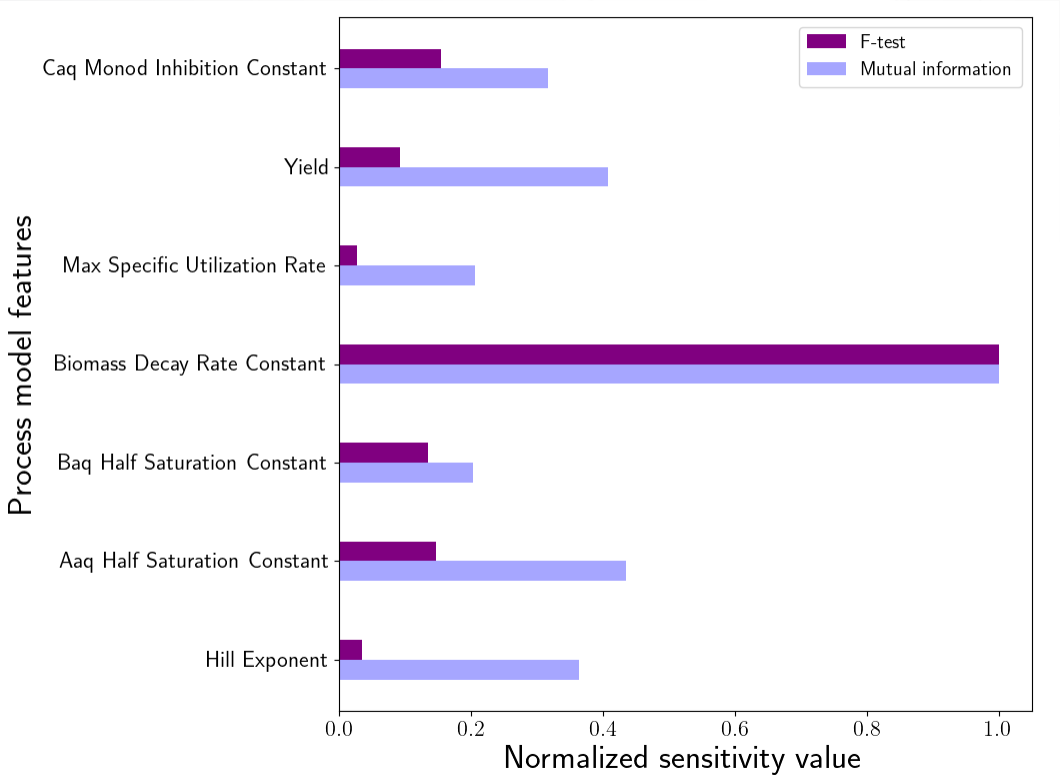


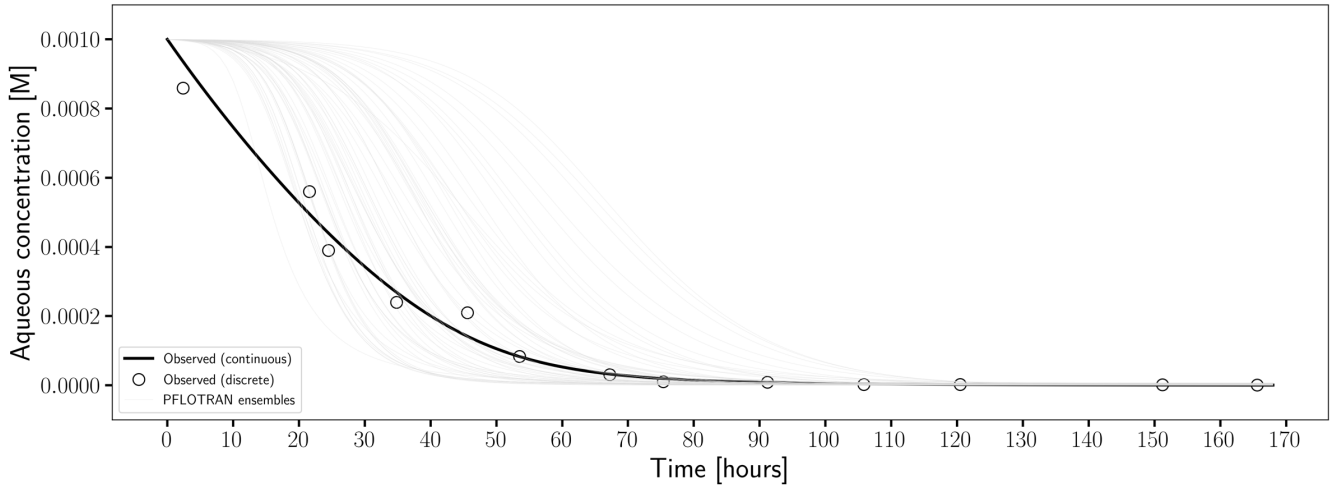
Step-3: Comprehensive sensitivity analysis (2/3)

- **Sensitivity analysis representative of observations**
  - F-test, MI, Random Forest, SHAPley values
  - Performance metrics
    - R^2, NSE, logNSE, KGE, mKGE, npKGE

Sensitivity (MI) for all realizations with logNSE-score > 0.5



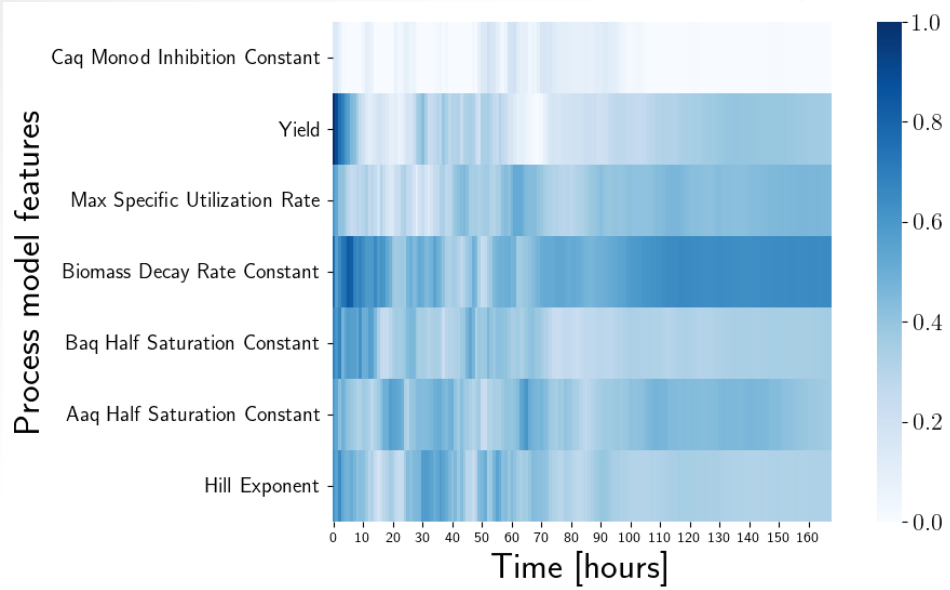




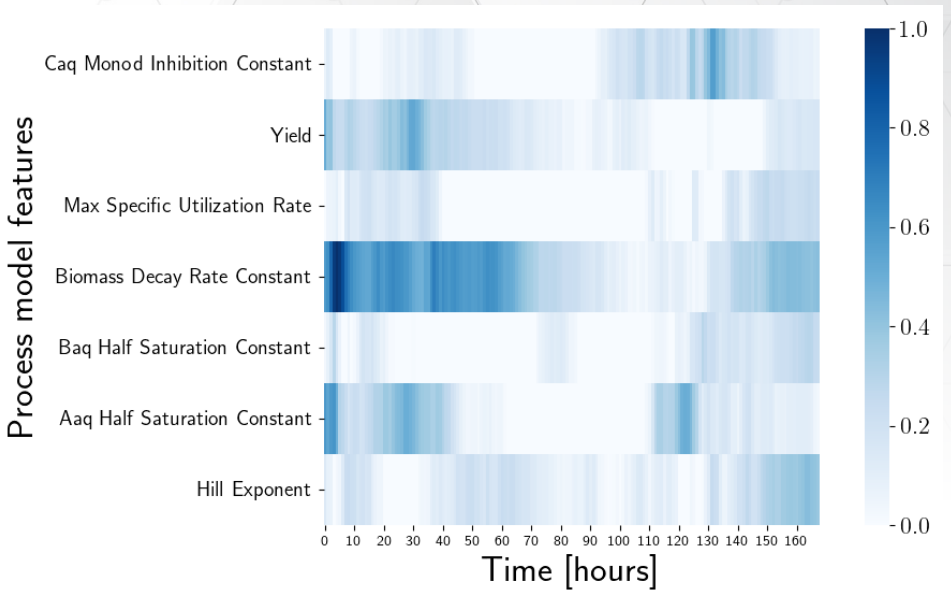
Step-3: Comprehensive sensitivity analysis (3/3)

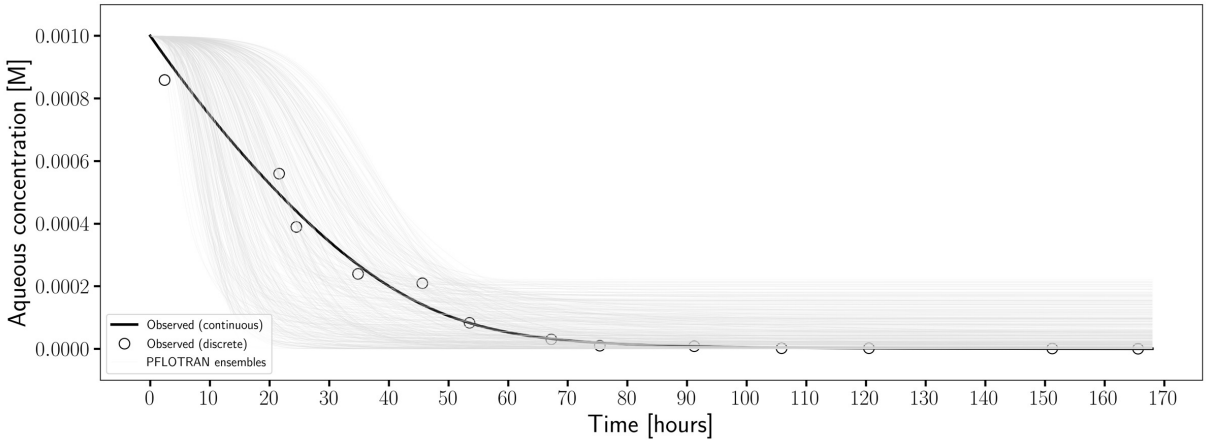
- **Local sensitivity analysis across realizations that are closer to the observational data based on R^2-score or logNSE > 0.5**

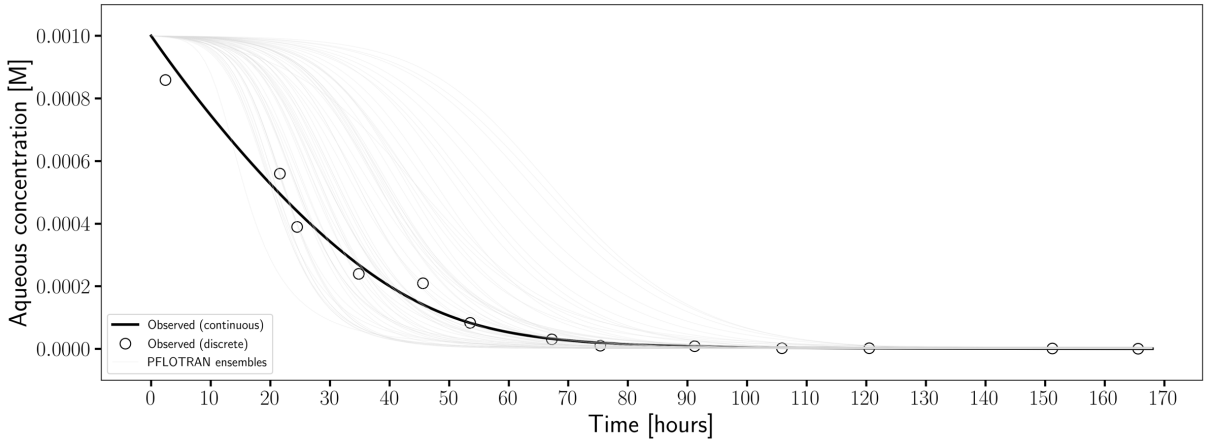
Sensitivity (MI) for all realz with R^2-score > 0.5



Sensitivity (MI) for all realz with logNSE-score > 0.5







Step-4 – ML model training and estimation of PFLOTRAN reaction network parameters

$$A_{aq} + 0.25B_{aq} \rightarrow 0.33C_{aq} + D_{aq},$$

Electron doner

Electron acceptor

1. Species time-series data

2. Data standardization

3. AI model – ML or DL + Hyperparameter tuning (on Tahoma)

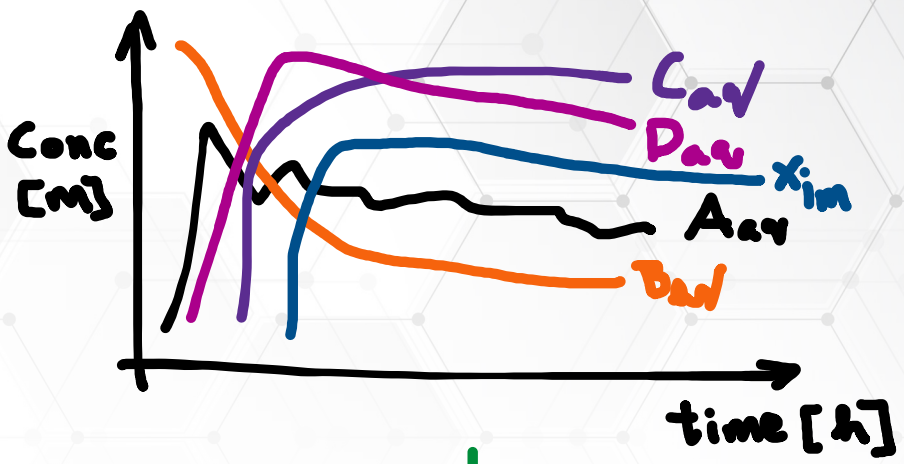
- Single task learning (STL)
  - Each parameter ➔ One model
  - Seven different models
- Multi-task learning (MTL)
  - All parameters ➔ One model

4. PFLOTRAN reaction network parameters

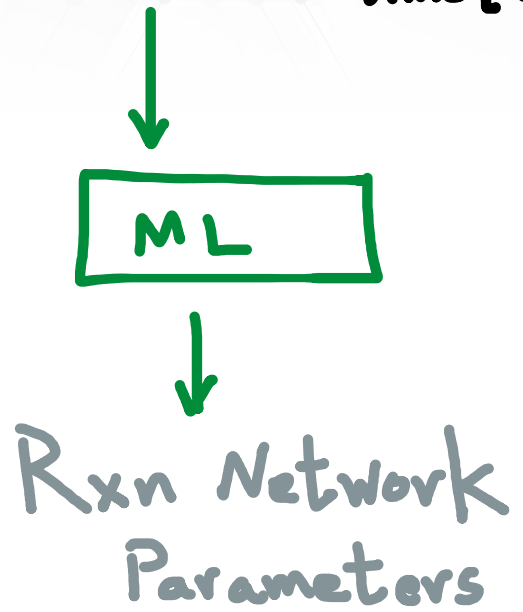
Step-4 – ML model training and estimation of PFLOTRAN reaction network parameters

Inputs and outputs of ML/DL model training

- Input to ML/DL model
  - Time-series data of chemical species



- Output to ML/DL model
  - Reaction network parameters



Step-4 – Component specifics

1. Species time-series data

- Aaq time-series
- Aaq + Baq + Caq + Daq + Xim

2. Data standardization

- Standard, MinMax, MaxAbs, Robust, Quantile Transformer, PowerTransformer Scalers
- Log10 transformation

3. AI model – ML or DL + Hyperparameter tuning (on Tahoma)

- Traditional ML – Random Forests, SVM
- Deep learning (DL) –
  - Deep neural networks (DNN)
  - Convolutional neural networks (CNN)

4. PFLOTRAN reaction network parameters

- Reaction sandbox parameters
  1. Max Specific Utilization Rate
  2. Aaq Half Saturation Constant
  3. Baq Half Saturation Constant
  4. Caq Monod Inhibition Constant
  5. Yield
  6. **Biomass Decay Rate Constant**
  7. Hill Exponent

Step-4 – Scenarios of ML model training

- Scenarios
  - Scenario-1: Full data (Aaq) + Multi-task learning
  - Scenario-2: GLUE data (Aaq) + Multi-task learning
  - Scenario-3: GLUE data (Aaq) + Single-task learning
  - Scenario-4: GLUE data (Aaq) + Single-task learning + reduced time-series steps
  - Scenario-5: GLUE data (Aaq) + Single-task learning + reduced time-series steps
  - **Scenario-6: GLUE data** (Aaq, Baq, Caq, Daq, Xim) **+ Single-task learning + reduced time-series steps**

GLUE → Generalized Likelihood Uncertainty Estimation

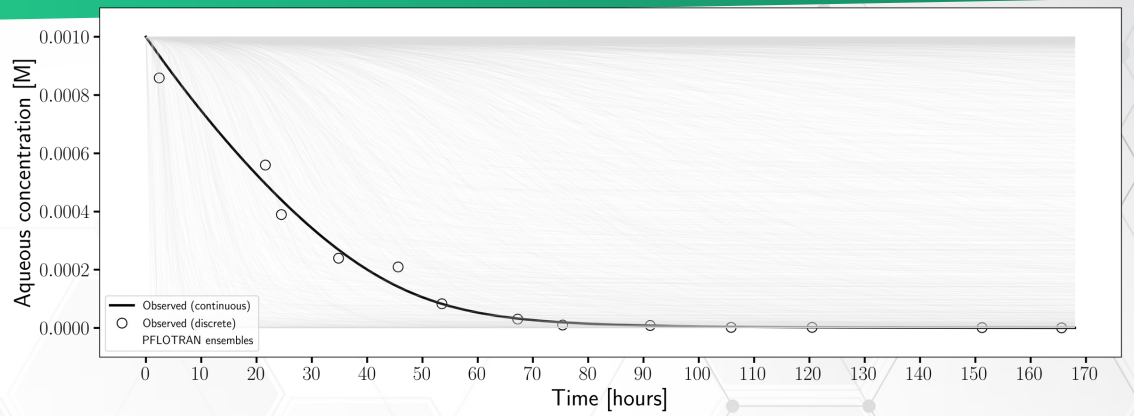
- A methodology to distill full data into actionable information for ML model training
- **A way to perform data worth analysis**

STL or Single-task learning → Train one ML model to predict each PFLOTRAN reaction network parameter

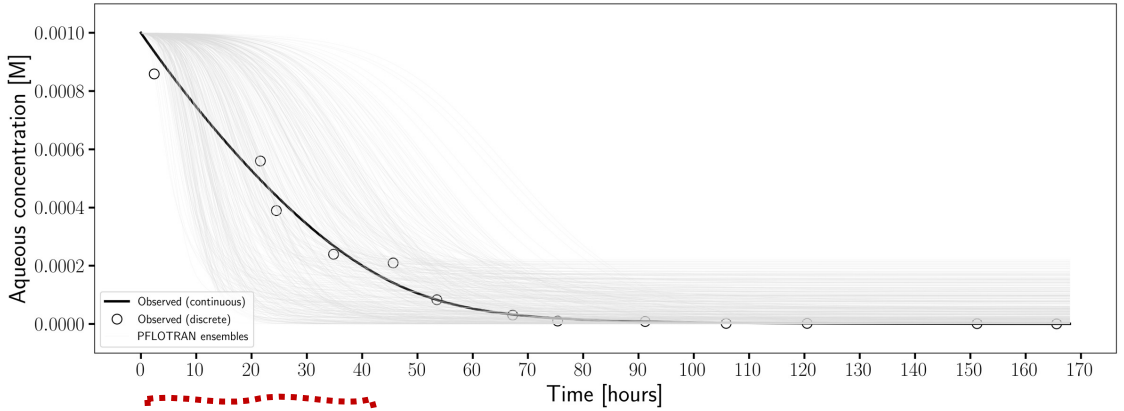
MTL or Multi-task learning → Train a single ML model that predicts all the PFLOTRAN reaction network parameters

Step-4 – Data worth analysis for ML model training

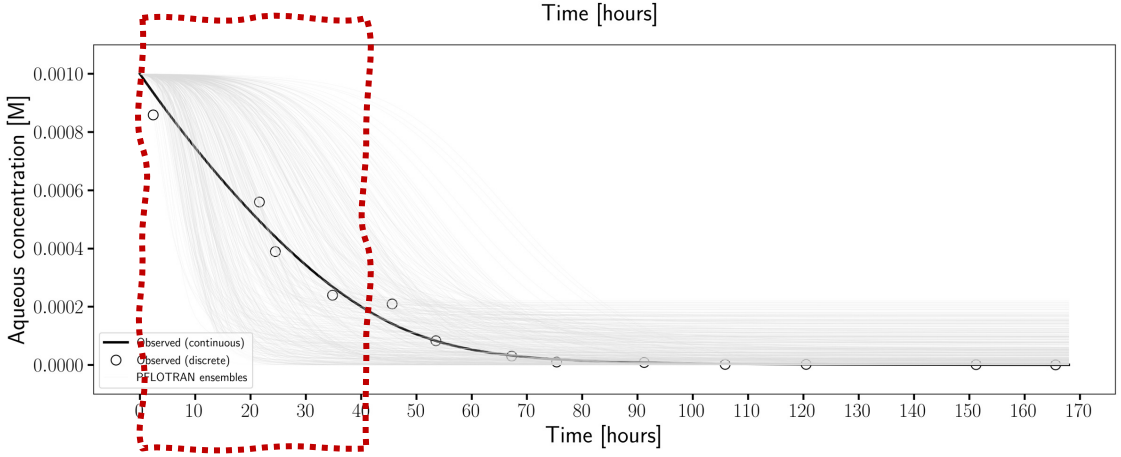
Full data – 15539 realz with 169 time-steps (7 days)



GLUE-based data – 489 realz with 169 time-steps (7 days)



GLUE-based data with reduced time-steps – 489 realz with 40 time-steps (1.5 days)

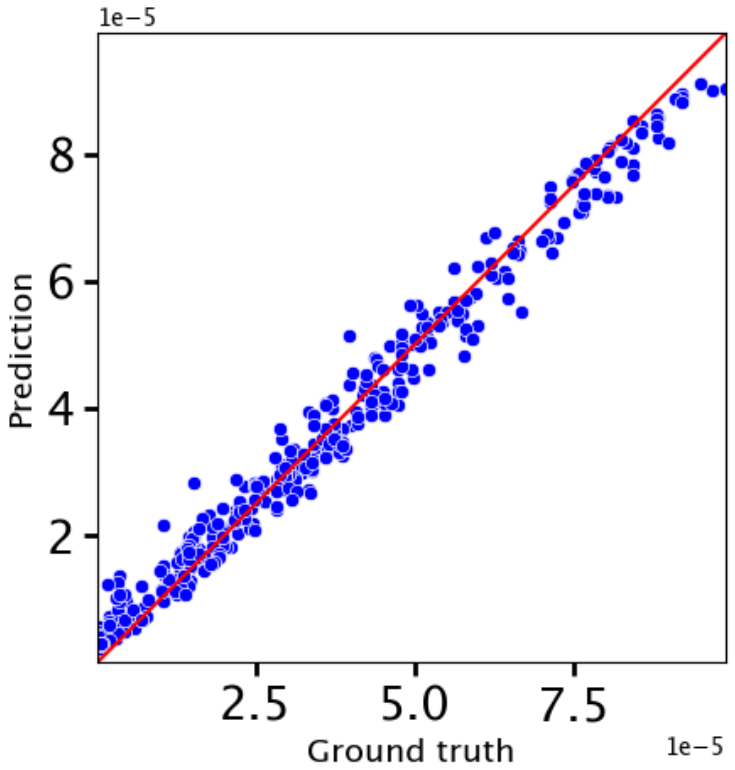


Step-5 – Preliminary results and main inference

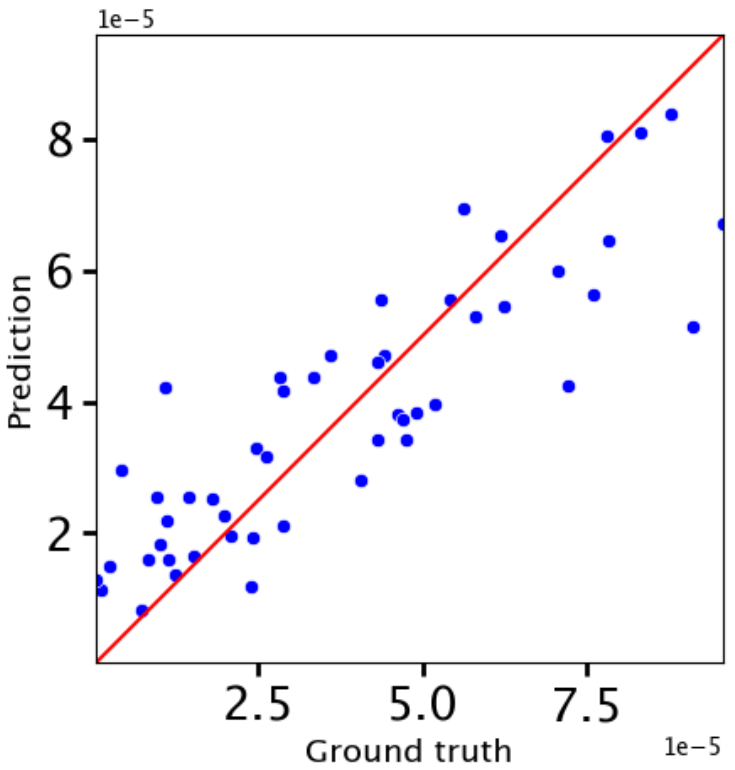
Scenario-6: ML + Single-task learning + reduced time-series steps (Inputs = Aaq + Baq + Caq + Daq + Xim)

- Biomass decay rate prediction

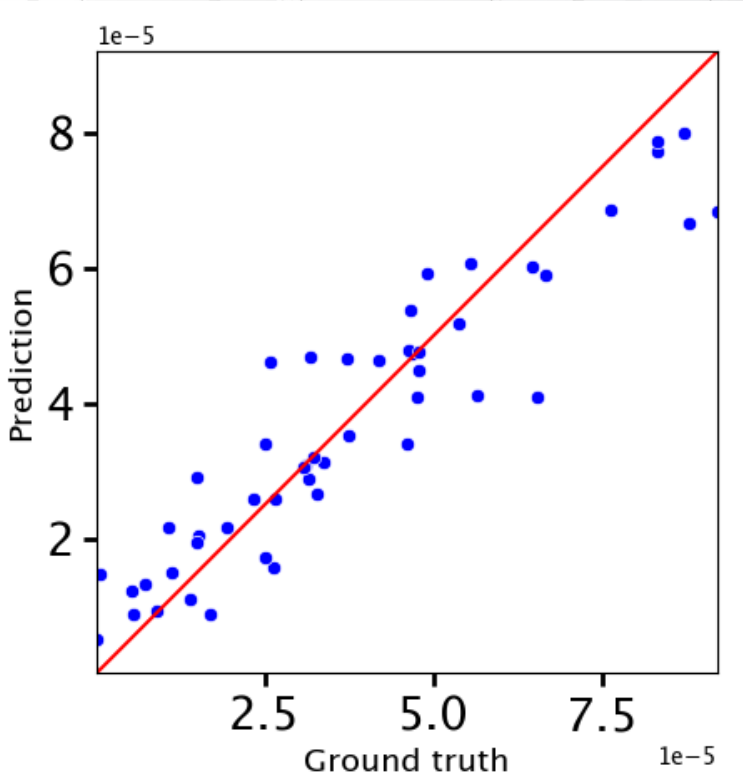
Train



Val



Testing –> R^2-score = 0.8

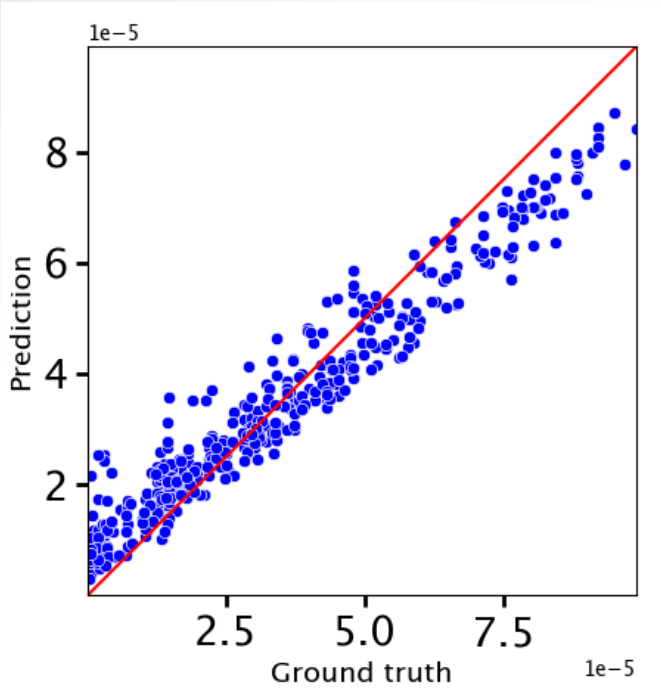


Step-5 – Results with only Aaq data

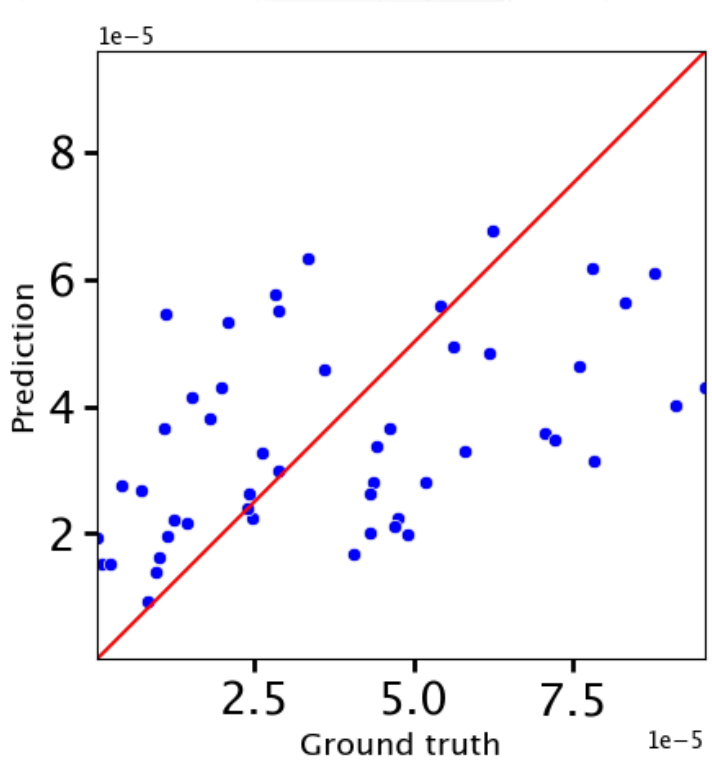
Scenario-6: ML + Single-task learning + reduced time-series steps (Inputs = Aaq only)

- Biomass decay rate prediction

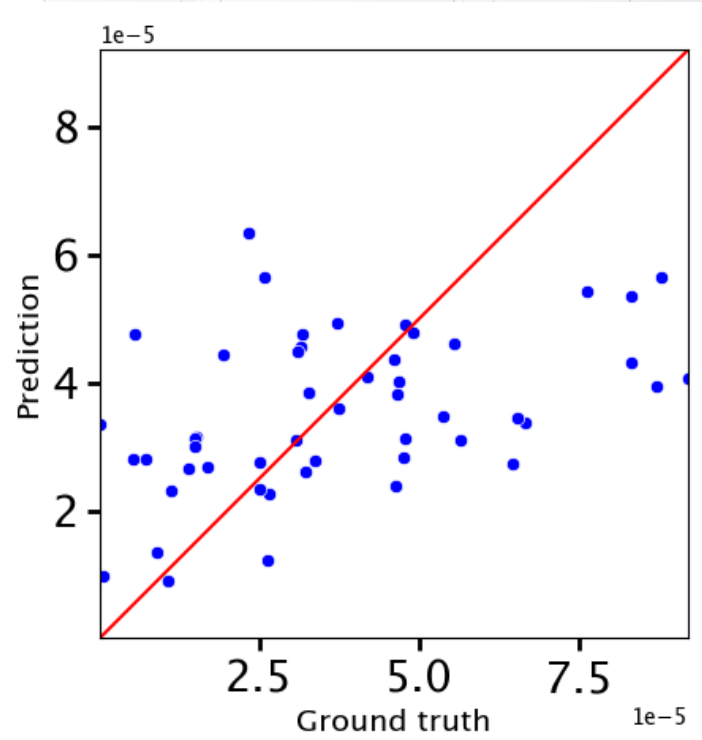
Train



Val



Testing –> R^2-score = 0.6



In-progress -- Hyperparameter tuning (DNNs and CNNs)

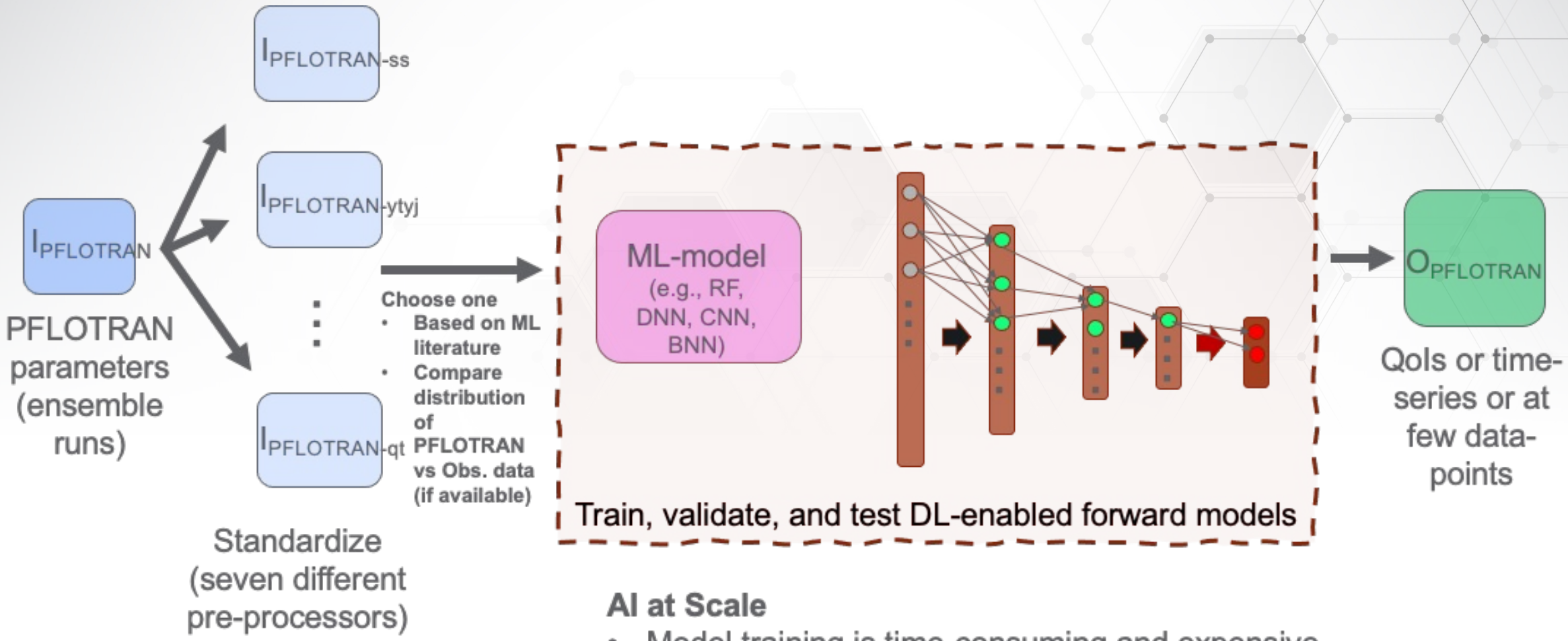
- DNNs – 21875 models trained for each scenario
  - Number of layers -- [1, 2, 3]
  - Neurons – [100, 50, 25]
  - Dropout values for reducing overfitting – [0.0, 0.1, 0.2, 0.3, 0.4]
  - Activation function (LeakyReLU) values -- [0.0, 0.1, 0.2, 0.3, 0.4]
  - Learning rate -- [1e-6, 1e-5, 1e-4, 1e-3, 1e-2]
  - Epochs -- [500, 1000, 1500, 2000, 2500]
  - Batch size -- [4, 8, 16, 32, 64]
- CNNs – 18750 models for each scenario
  - Number of layers -- [1]
  - Filter sizes – [256, 128, 64, 32, 16]
  - Kernel sizes -- [32, 16, 8, 4, 2]
  - Dropout values for reducing overfitting – [0.0, 0.1, 0.2, 0.3, 0.4]
  - Learning rate -- [1e-6, 1e-5, 1e-4, 1e-3, 1e-2]
  - Epochs -- [50, 100, 200, 300, 400, 500]
  - Batch size -- [4, 8, 16, 32, 64]

In-progress -- Popular pre-processors for ML model development

- Standard scaler
- Max-absolute scaler
- Min-Max Scaler
- Robust Scaler
- Quantile Transformer
  - Normal distribution
  - Uniform distribution
- Power Transformer

- Why need them?
  - Each pre-processor has it own advantage
  - Pre-processor has considerable effect on ML-model training/predictions/inference
  - We need pre-processor that do well on data with outliers
    - (due to process model structural errors)

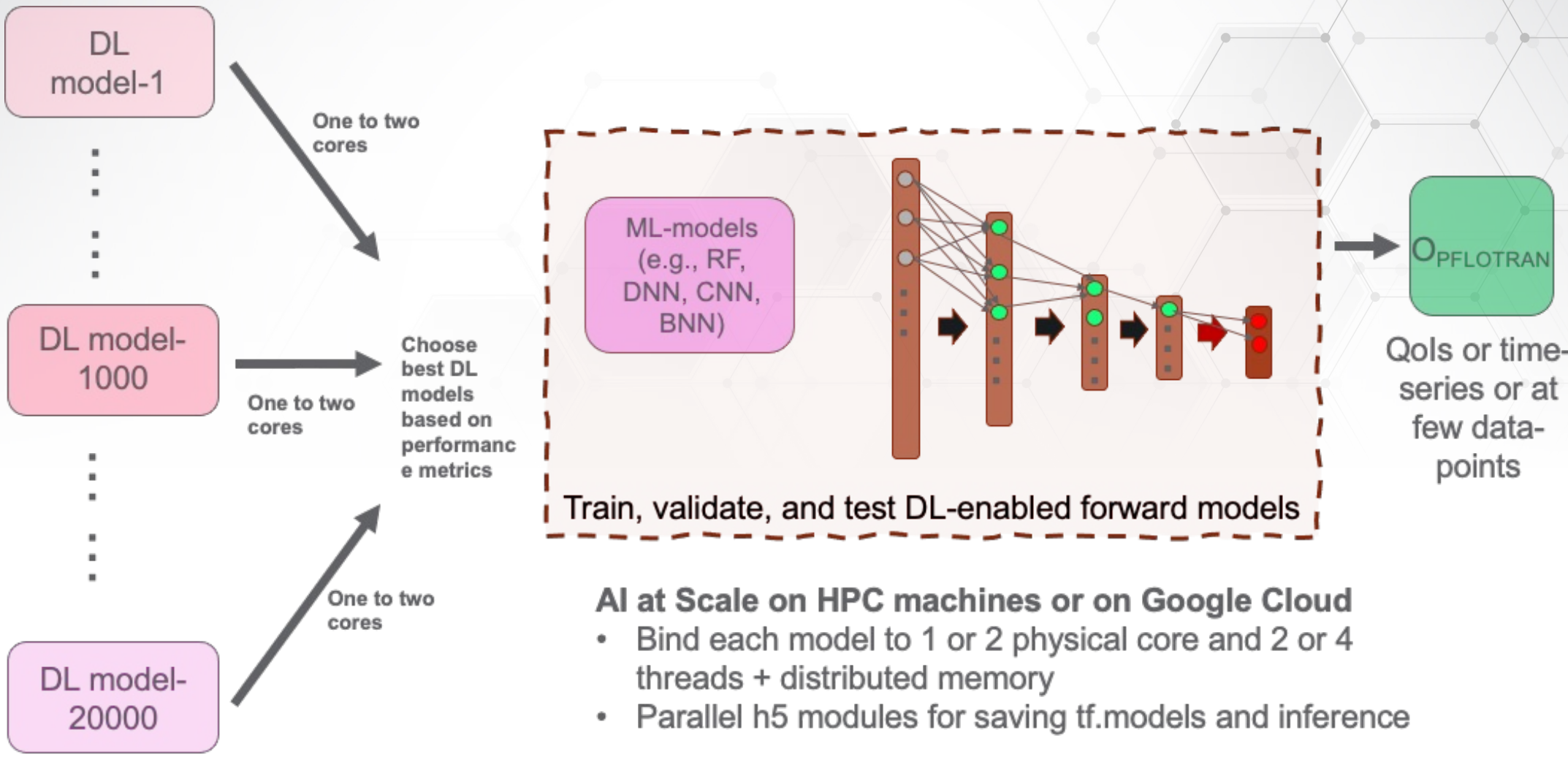
In-progress: Train, validate, and test ML models at scale (1/2)



AI at Scale

- Model training is time-consuming and expensive
- Model training is embarrassingly parallel
- Hyperparameter tuning using grid search

In-progress: Train, validate, and test ML models at scale (2/2)



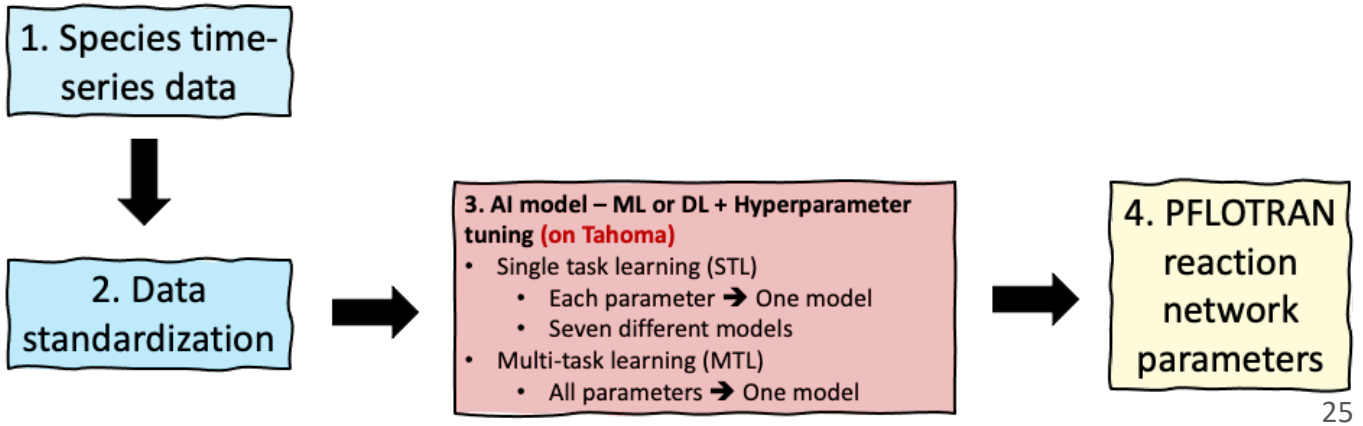
AI at Scale on HPC machines or on Google Cloud

- Bind each model to 1 or 2 physical core and 2 or 4 threads + distributed memory
- Parallel h5 modules for saving tf.models and inference

Conclusions, outcomes, and Impact

What have we achieved? How the results impact PFLOTRAN users?

- Before this project there was no AI-enabled ModEx workflow for PFLOTRAN users.
- AI shows promise in parameter estimation and inverse modeling based on performance metrics (good R^2-scores)
- Informative data (e.g., GLUE, adding other species) can easily improve AI model performance
- PFLOTRAN users can use this workflow or components of it in their research (e.g., data analysis, sensitivity analysis)



Our next steps: Improving the AI/ML workflows

Key takeaways:

- Preliminary results show promise in AI/ML models but extensive hyperparameter tuning is necessary
- AI-enabled ModEx: A scalable AI workflow to efficiently calibrate PFLOTRAN process models
  - Manuscript in development with team members and collaborators – Glenn Hammond, Cheng Shi, Satish Karra, and Emily Graham

Once the workflow is fully tested and validated, Python scripts will be made available and open-sourced at: https://github.com/maruti-iitm/AI_ModEx_PFLOTRAN.git

"This research was performed on a project award from the Environmental Molecular Sciences Laboratory, a DOE Office of Science User Facility sponsored by the Biological and Environmental Research program under Contract No. DE-AC05-76RL01830."

Questions?