人工智慧與多尺度水文資訊研究室

從參數校正到參數學習

From calibration to parameter learning

機器學習於地球科學系統之應用

Dr. Wen-Ping Tsai (蔡文柄)

2025.05.05

ABOUT ME

Education

Degree	University
Ph.D. (2014)	Dept. of Bioenvironmental Systems Engineering, National Taiwan University, Taiwan

Academic Experience

- **National Cheng Kung University, Dept. of Hydraulic and Ocean Engineering – Assistant Professor(2022/8~)**
- Pennsylvania State University, Dep. of Civil and Environmental Engineering – Postdoctoral Fellow (2017/4-2022/4)
- National Taiwan University, Dep. of Bioenvironmental Systems Engineering – Postdoctoral Fellow (2014/7-2017/4)
- University of Illinois at Urbana-Champaign (UIUC), Illinois State Geological Survey (ISGS) - Visiting Scholar (2012/6-2013/6)

01

Research in TAIWAN

03

Integrating process-based models and ML

05

Projects

02

PENN STATE UNIV

04

Harnessing the scaling effects of big data in geoscientific modeling

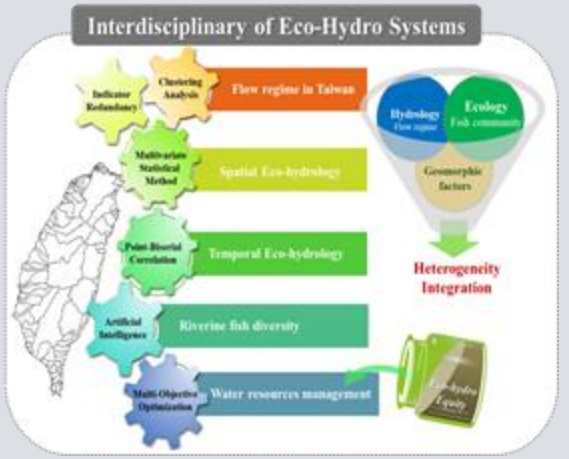
AIMH

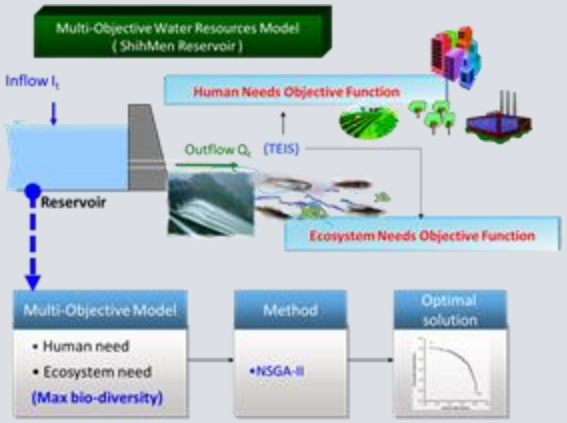
PART.01

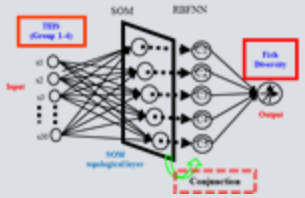
Research in Taiwan

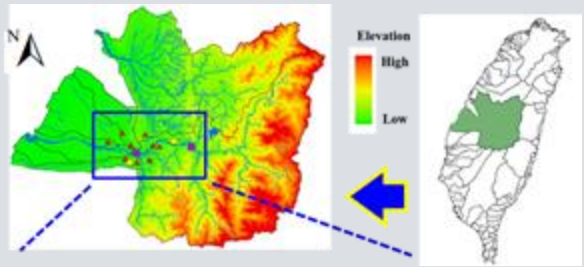
AIMH

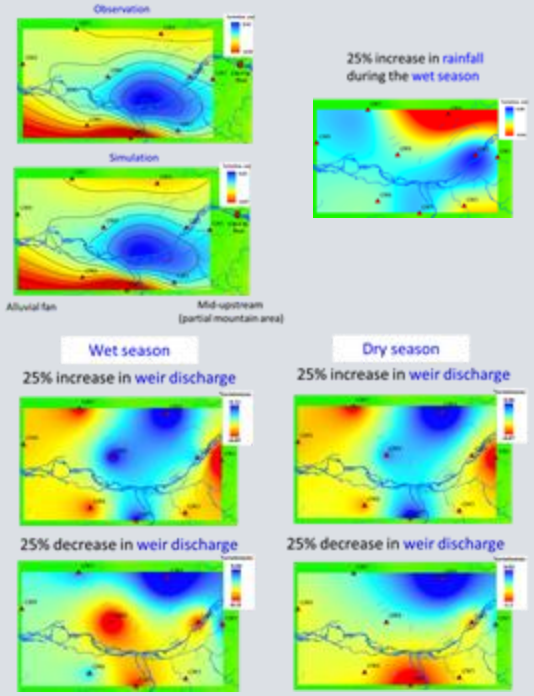
Research in Taiwan

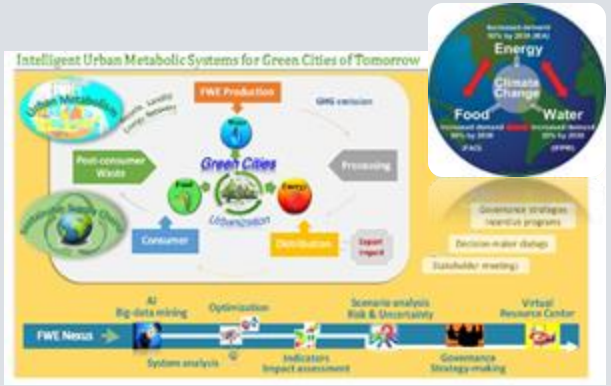




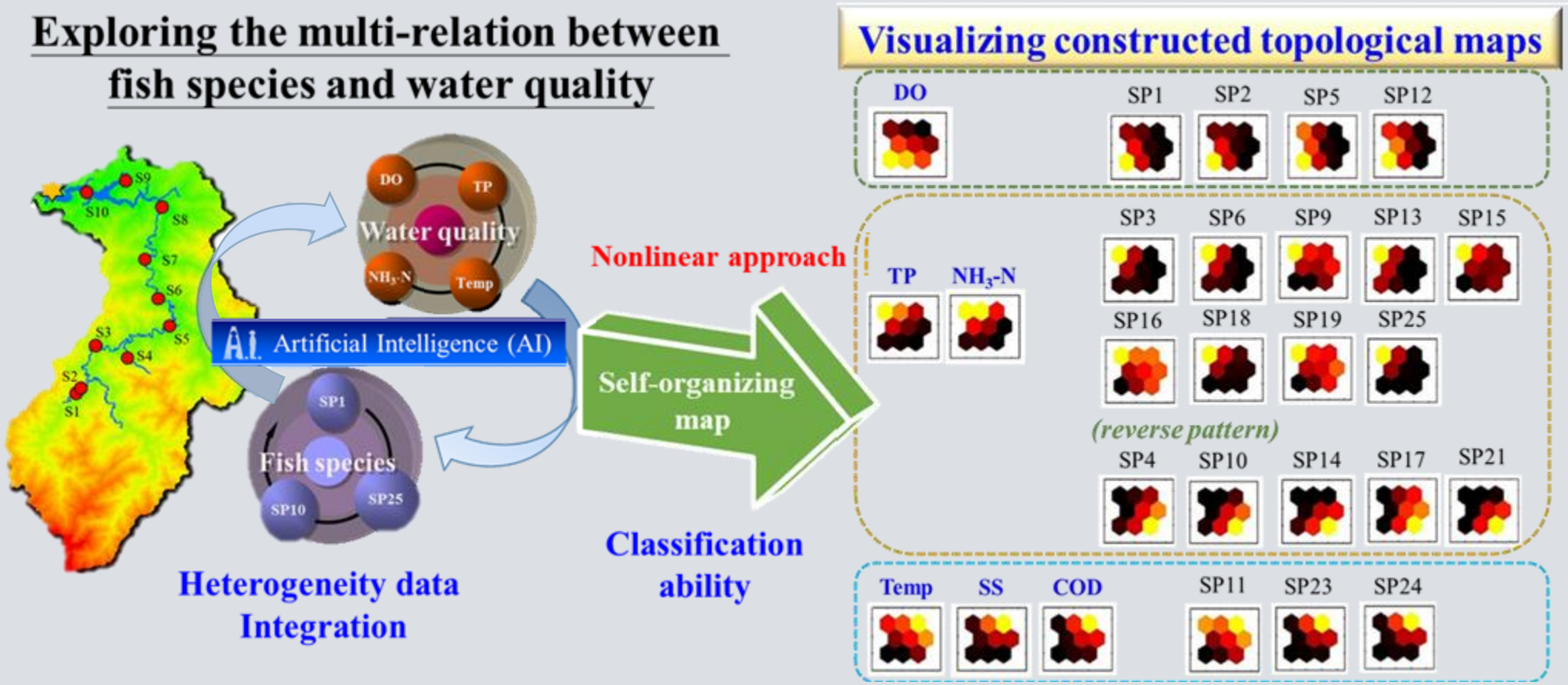




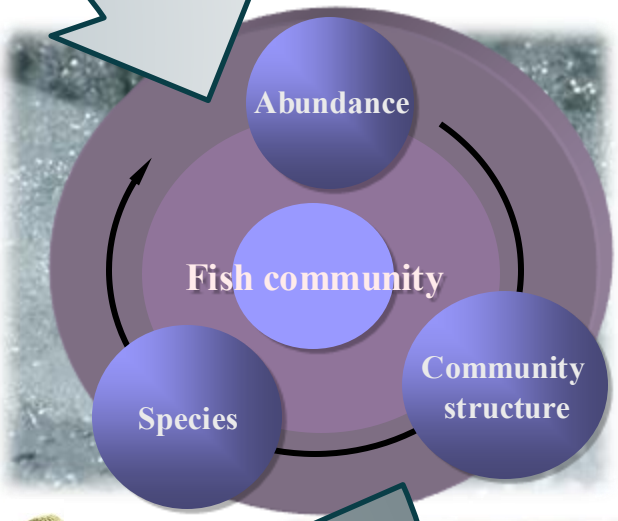


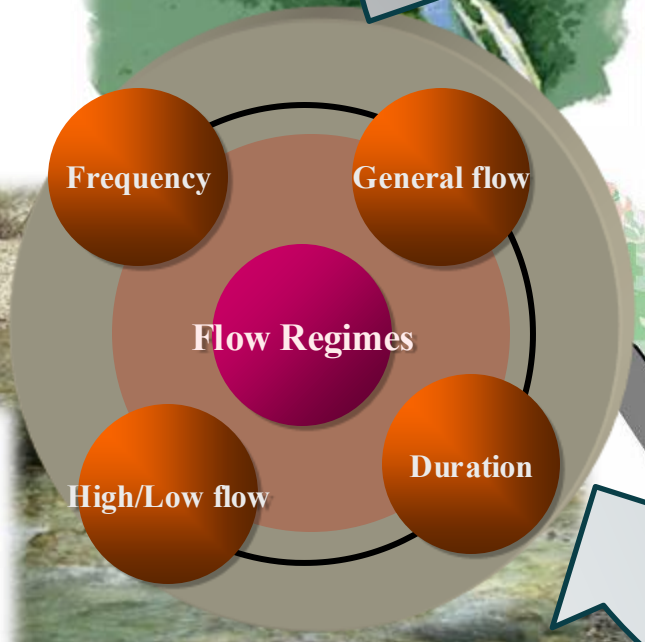


A data-mining framework for exploring the multi-relation between fish species and water quality through self-organizing map



The ecological community of the flow regime has become stable after long-term evolution.

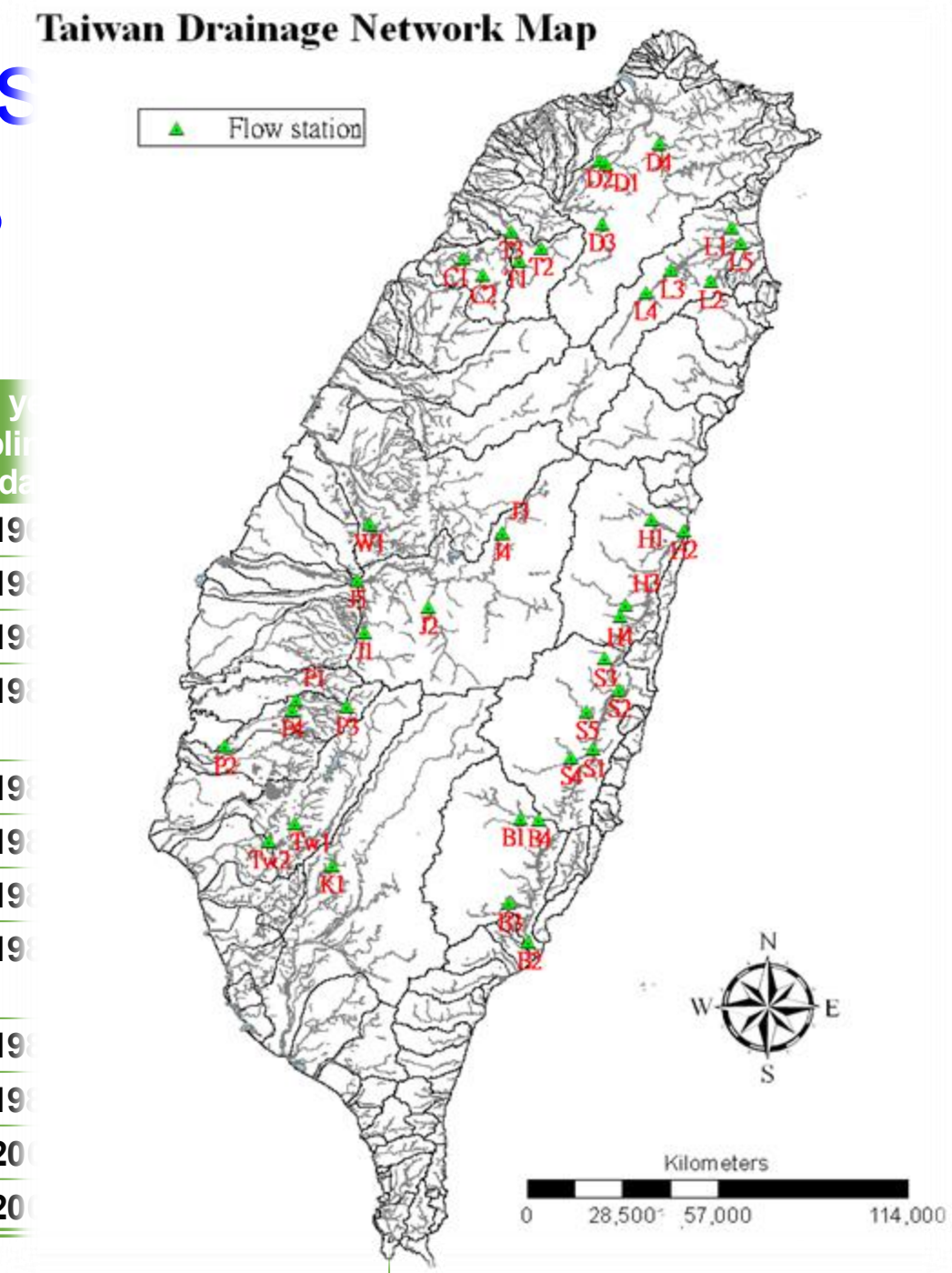




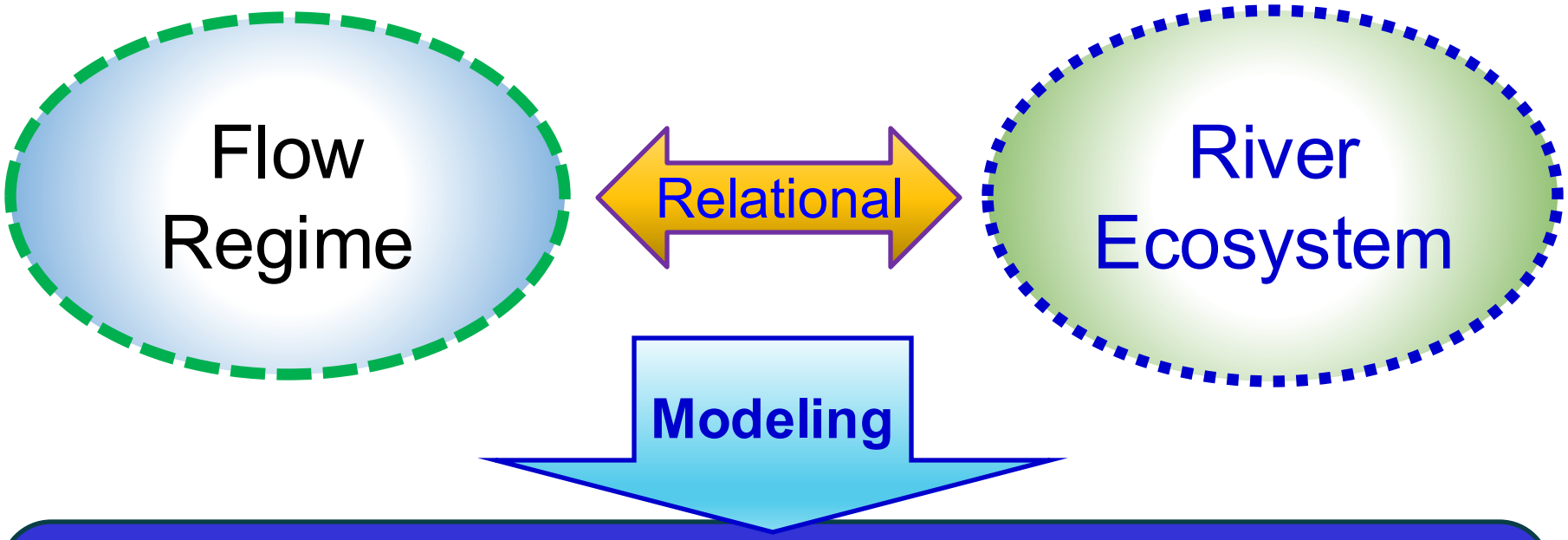
A problem in developing relationships between hydrologic statistics and sampled communities is that the presence and abundance of organisms at the time of the sampling are related as much antecedent conditions as to any conditions measured at the time of sampling.

Flow gauge stations
Fish sampling sites

Basins	Flow gauge stations & Fish sampling sites	Collecting y fish sampli da
Danshui River Basin	4 (D1~D4)	1981-1998 (19
Touchien River Basin	3 (T1~T3)	2004-2006 (19
Lanyang River Basin	5 (L1~L5)	2000-2004 (19
Siouguluan River Basin	5 (S1~S5)	2004-2006 (19
Beinan River Basin	4 (B1~B4)	2002-2003 (19
Hualien River Basin	4 (H1~H4)	2002-2003 (19
PaChang River Basin	4 (P1~P4)	2004-2005 (19
ChungKang River Basin	2 (C1, C2)	2006-2007 (19
Kaoping River Basin	1 (K1)	2005-2006 (19
Tsengwen River Basin	2 (Tw1, Tw2)	2005-2006 (19
Jhuoshuei River Basin	5 (J1~J5)	2005 (1986-20
Wu River Basin	1 (W1)	2005 (1986-20
Total	40	



Research Objectives



1. Find the relationship between flow regime and fish biodiversity.
2. Estimate the biodiversity according to the flow regime.
3. Provide a basis for river restoration

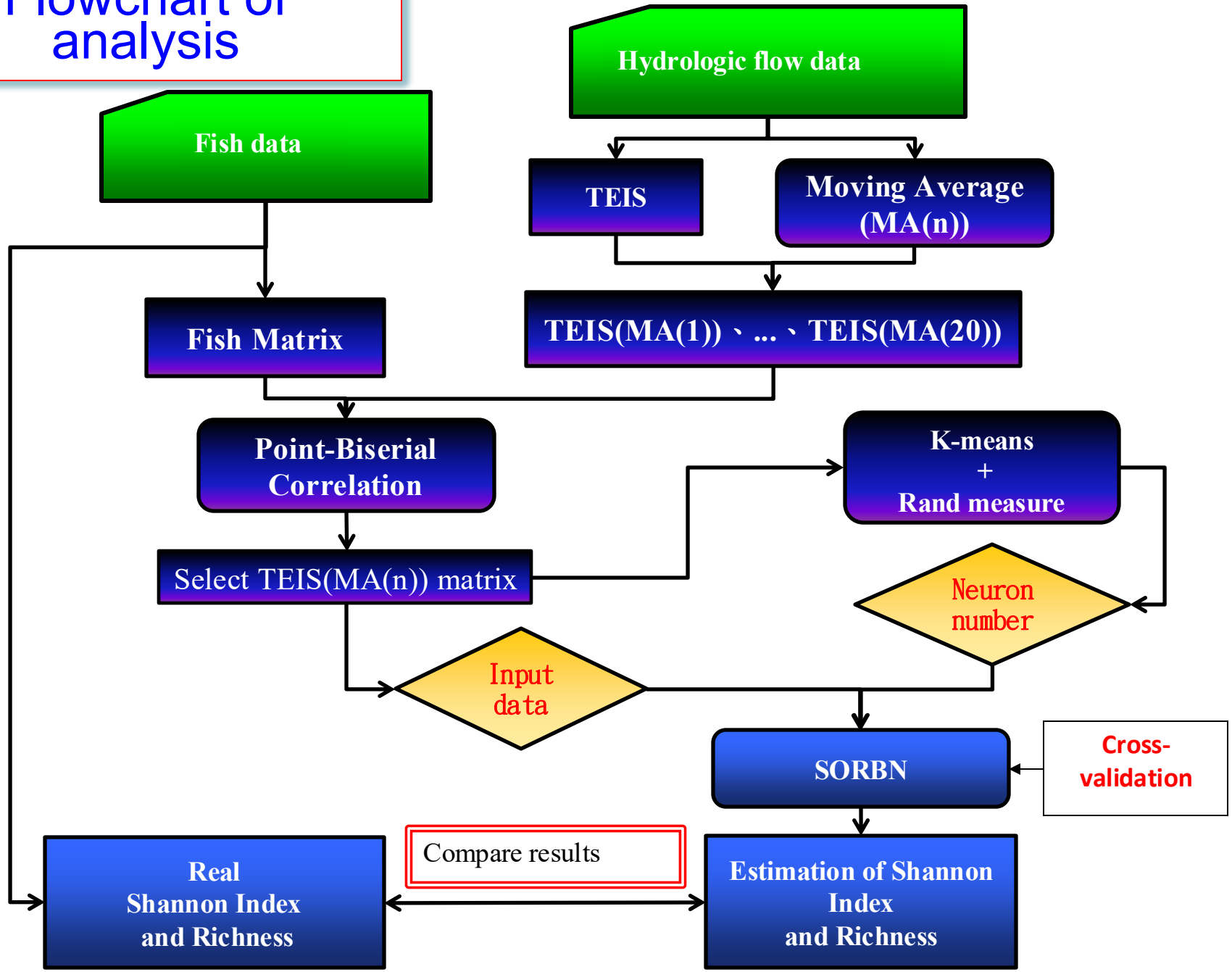
Before modeling

- Two questions :

1. How to choose the input data?

2. How to decide the number of neurons?

Flowchart of analysis



Input data

Fish Data Matrix

Fish 1	Fish 2	Fish 3	…	Fish 57
1	0	1	…	0
0	0	0	…	1
1	1	0	…	0
0	0	1	…	1
0	1	1	…	0
:	:	:	:	:

Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
:

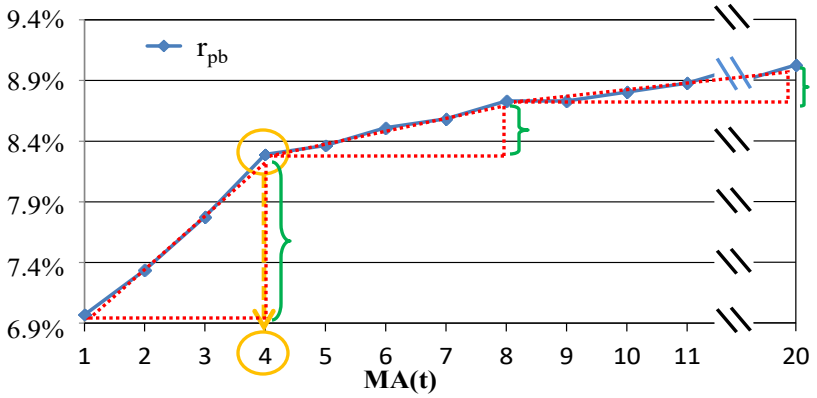
TEIS Matrix of MA(t)

TEIS 1	TEIS 2	TEIS 3	…	TEIS 66
3.2	64.6	1.2	…	19.2
4.1	51.9	2.4	…	51.2
3.4	4.7	13.5	…	9.2
51.7	9.5	7.7	…	20.5
13.6	1.1	53.1	…	51.1
:	:	:	:	:

Point-Biserial Correlation

rpb Matrix

	TEIS 1	TEIS 2	TEIS 3	…	TEIS 66
Fish 1	$r_{pb\ 1,1}$	$r_{pb\ 1,2}$	$r_{pb\ 1,3}$	…	$r_{pb\ 1,66}$
Fish 2	$r_{pb\ 2,1}$	$r_{pb\ 2,2}$	$r_{pb\ 2,3}$	…	$r_{pb\ 2,66}$
Fish 3	$r_{pb\ 3,1}$	$r_{pb\ 3,2}$	$r_{pb\ 3,3}$	…	$r_{pb\ 3,66}$
:	:	:	:	…	:
Fish 57	$r_{pb\ 57,1}$	$r_{pb\ 57,2}$	$r_{pb\ 57,3}$	…	$r_{pb\ 57,66}$



Neuron number

Rand measure

a:
the number of pairs of elements that are in the same set in C and in the same set in K

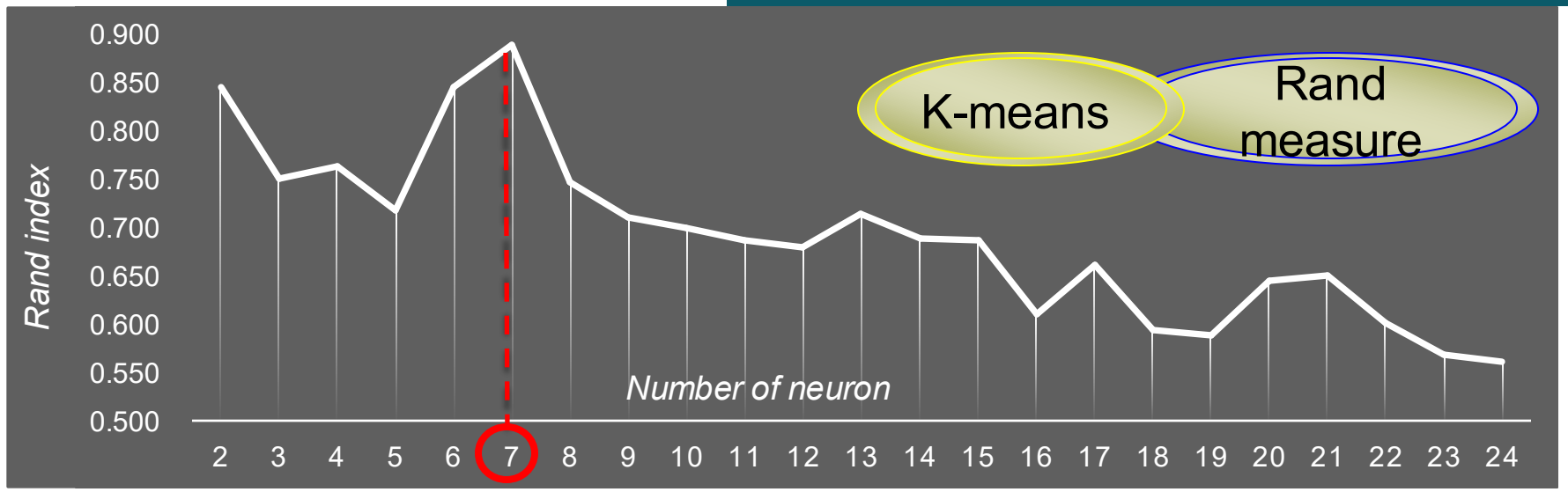
b:
the number of pairs of elements that are in different sets in C and in different sets in K

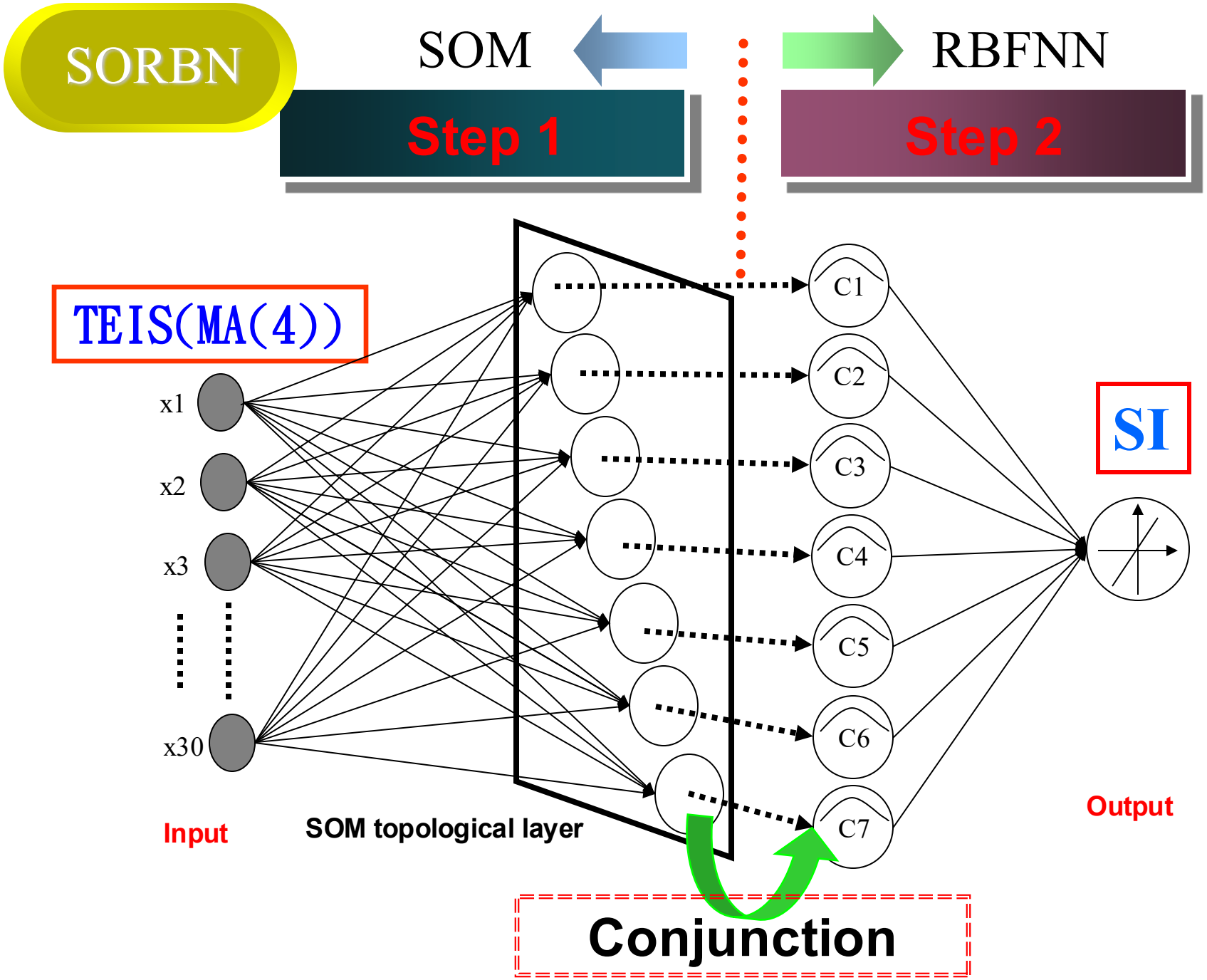
c:
the number of pairs of elements that are in the same set in C and in different sets in K

d:
the number of pairs of elements that are in different sets in C and in the same set in K

$$x=a+b+c+d$$

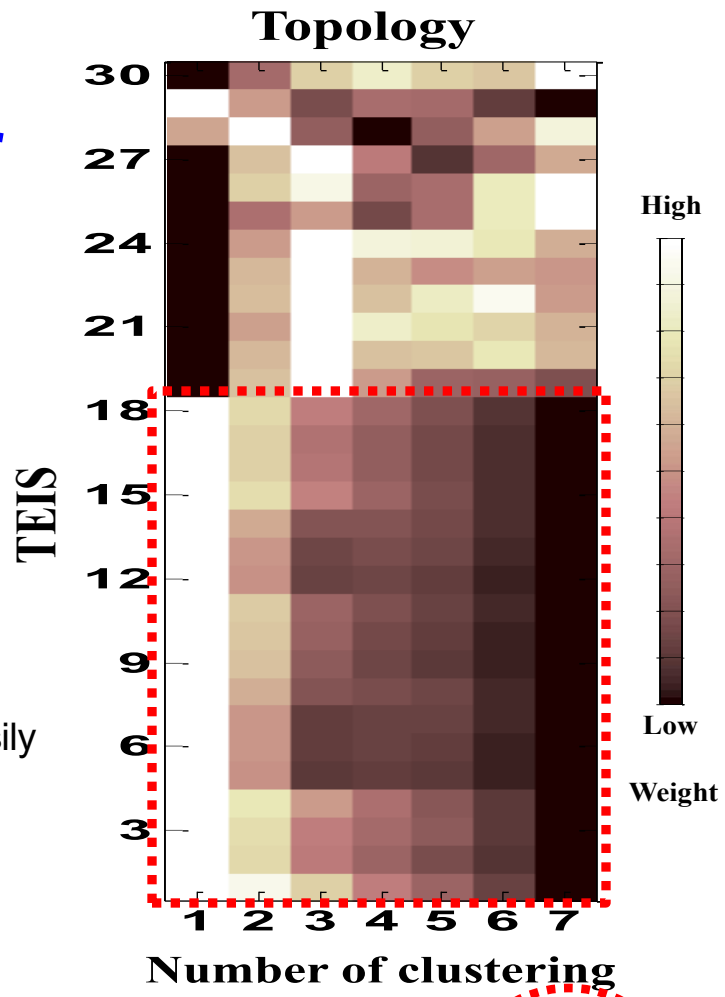
$$\text{Rand Index}=(a+b)/x$$




SORBN
SOM
RBFNN
Step 1
Step 2
TEIS(MA(4))
x1
x2
x3
x30
C1
C2
C3
C4
C5
C6
C7
SI
Input
SOM topological layer
Output
Conjunction

Topological information of SOM

- The first 18 TEIS have consistent characteristics and dominate the classification of datasets.

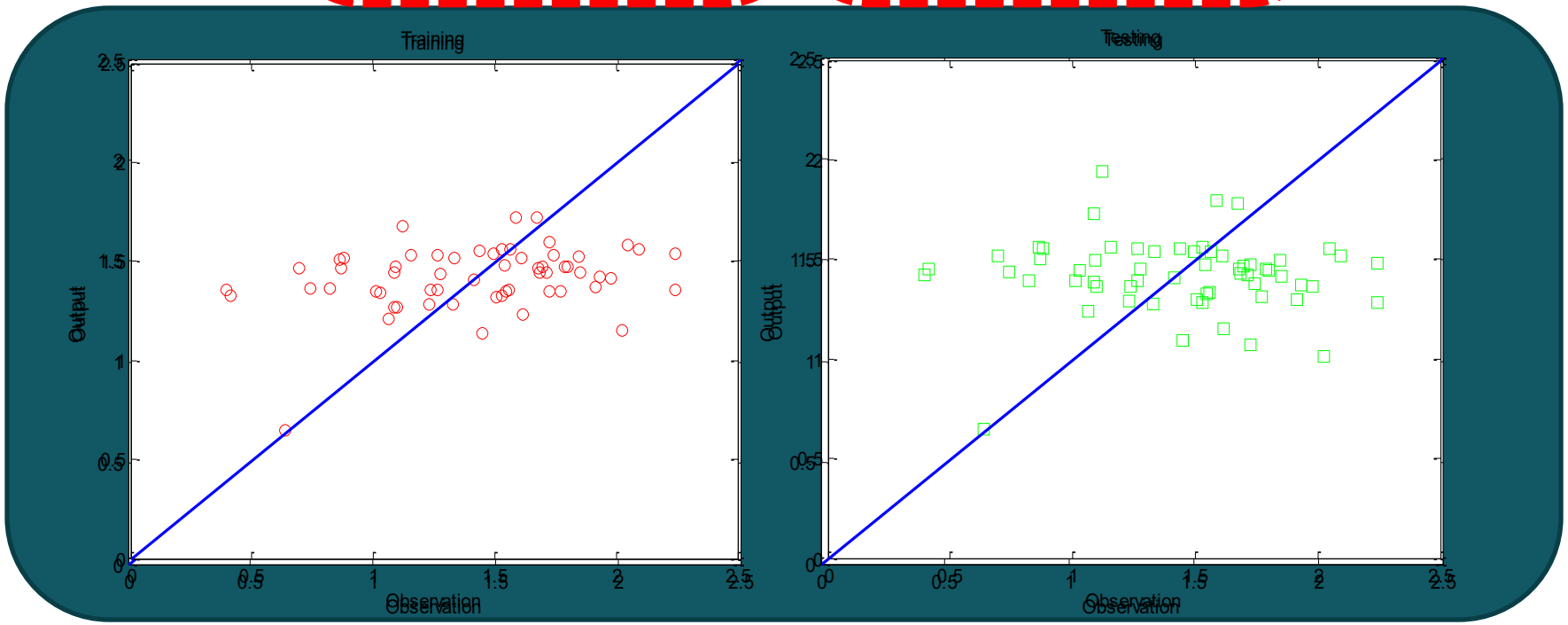


- The stations with the highest and the lowest SI values can be easily identified by the SOM.
- Clusters 5-7 have relatively wide ranges and large variances of SI values.

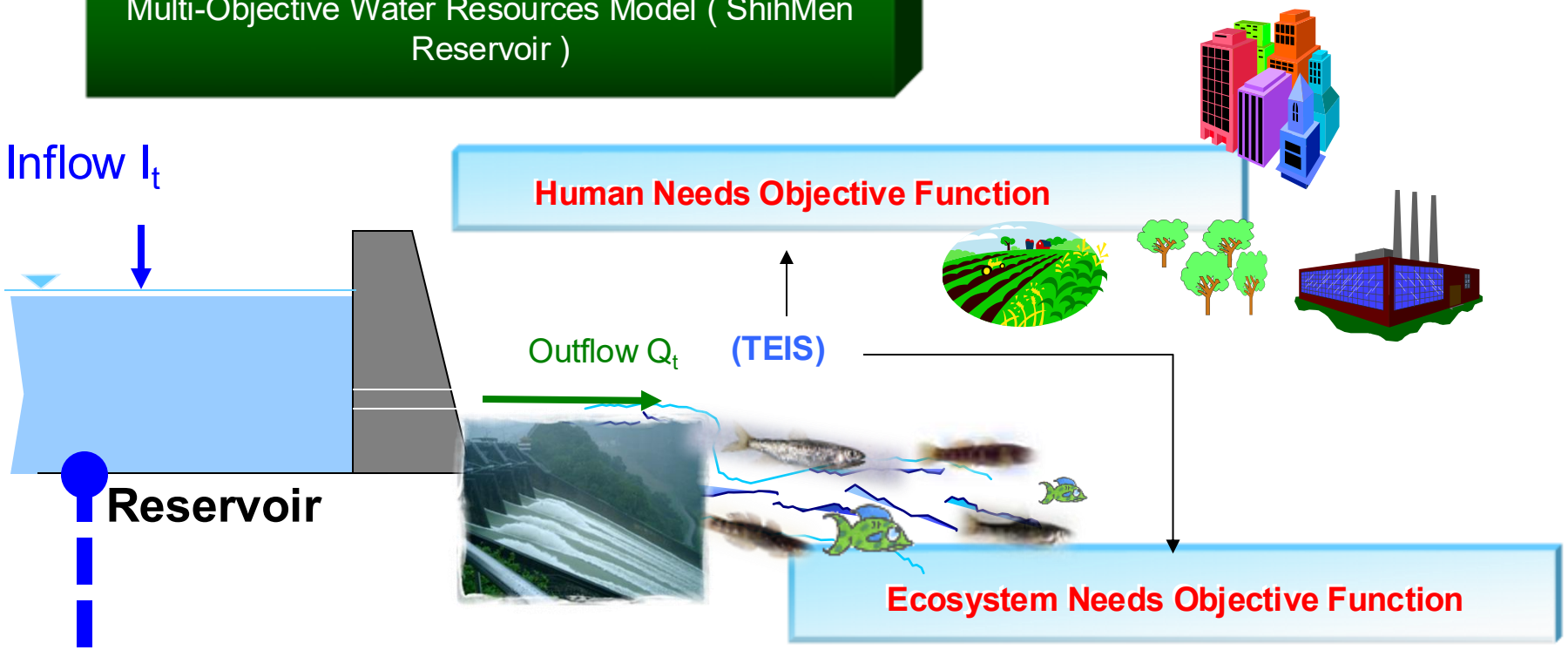
Cluster	1	2	3	4	5	6	7
Range of SI	1.06-1.74	1.26-1.72	1.41-2.24	1.1-2.02	0.41-1.85	0.82-2.10	0.39-1.80
Variance of SI	0.054	0.11	0.11	0.13	0.18	0.22	0.16

The results of Shannon index

MA 1				MA 4			
Training		Testing		Training		Testing	
RMSE	CC	RMSE	CC	RMSE	CC	RMSE	CC
0.42	0.28	0.48	-0.03	0.33	0.64	0.38	0.52



Multi-Objective Water Resources Model (ShihMen Reservoir)



Multi-Objective Model

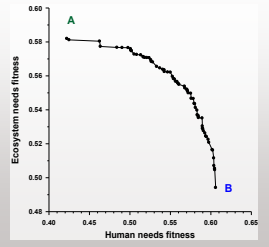
- Human need
- Ecosystem need

(Max bio-diversity)

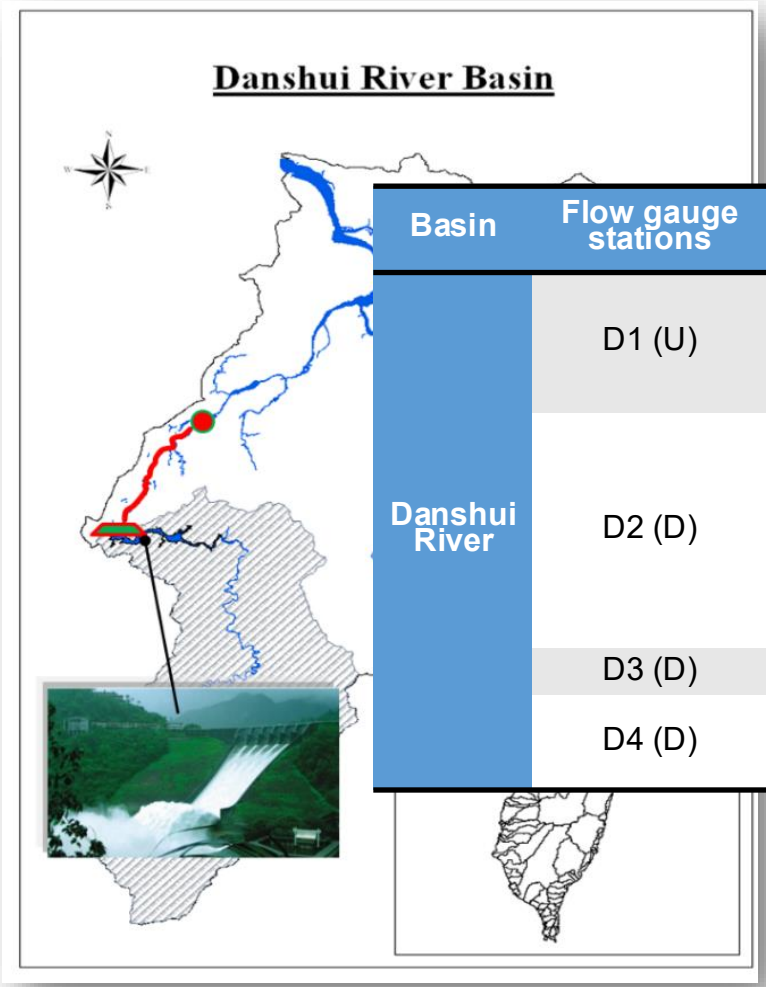
Method

- NSGA-II

Optimal solution

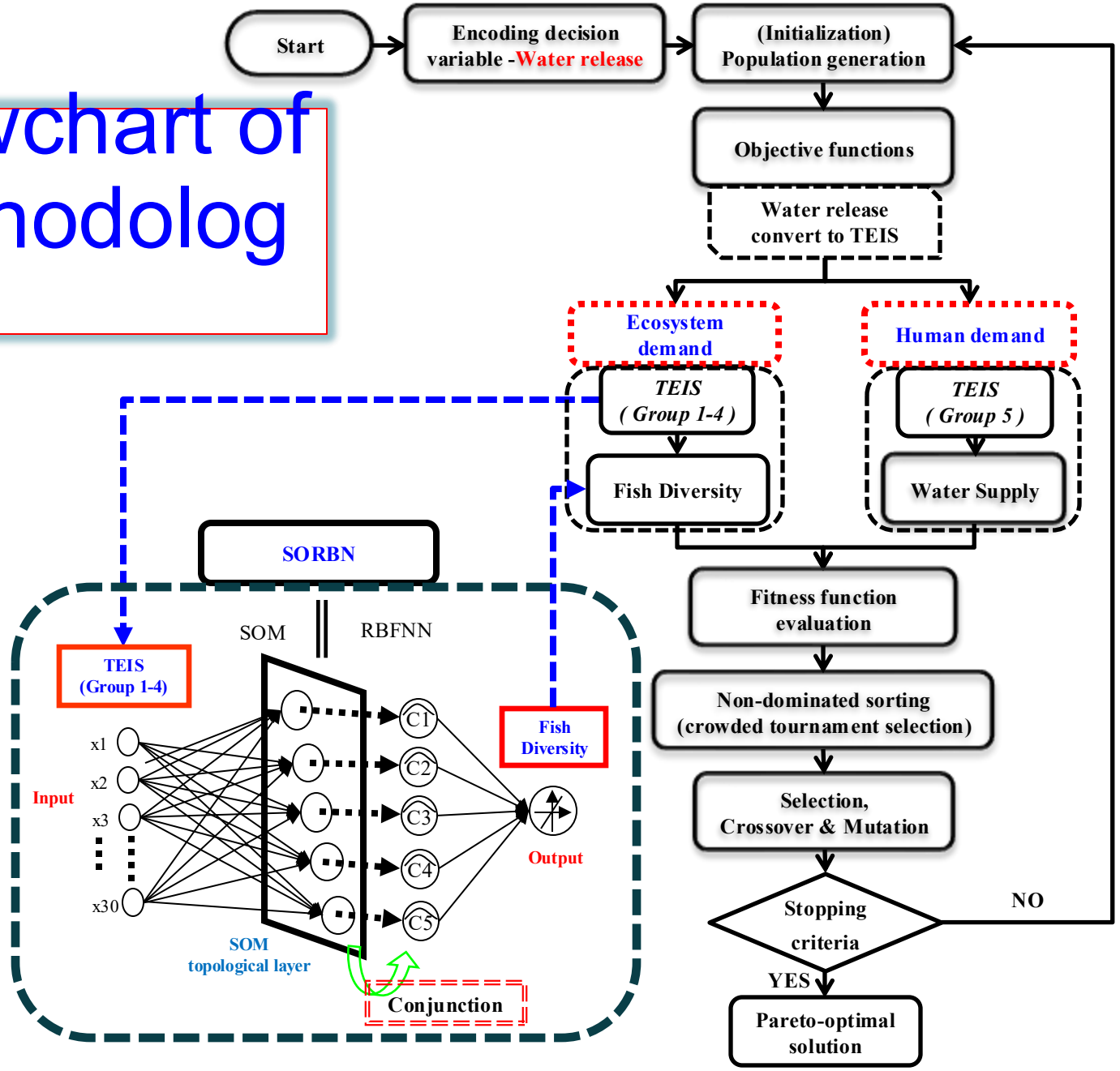


Study area

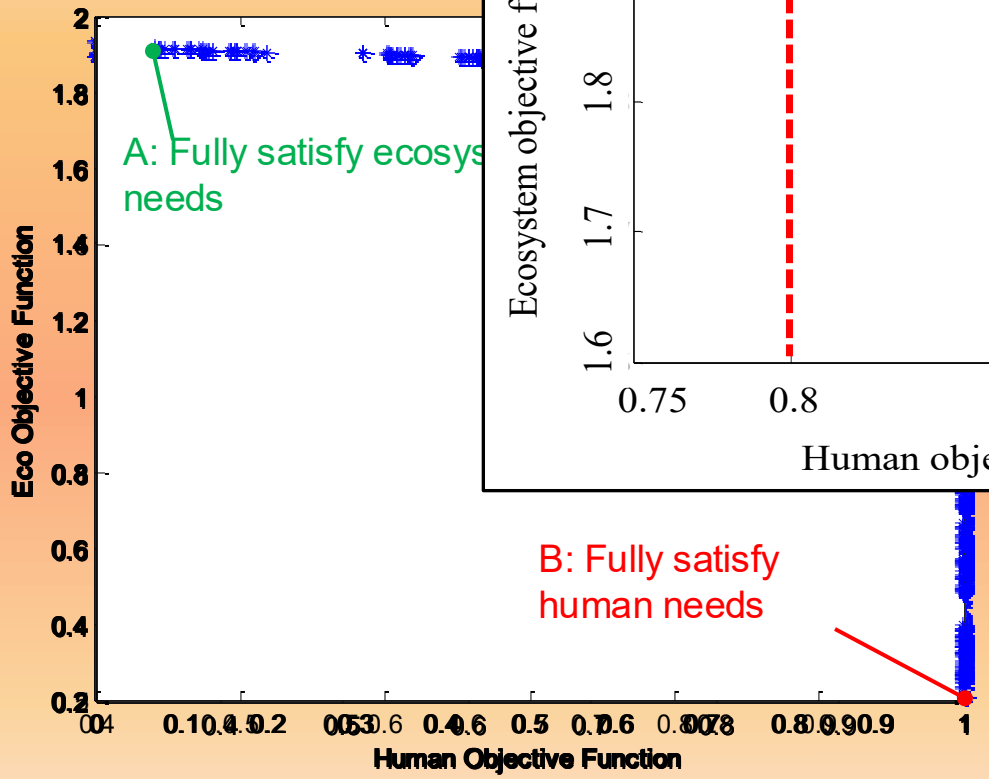


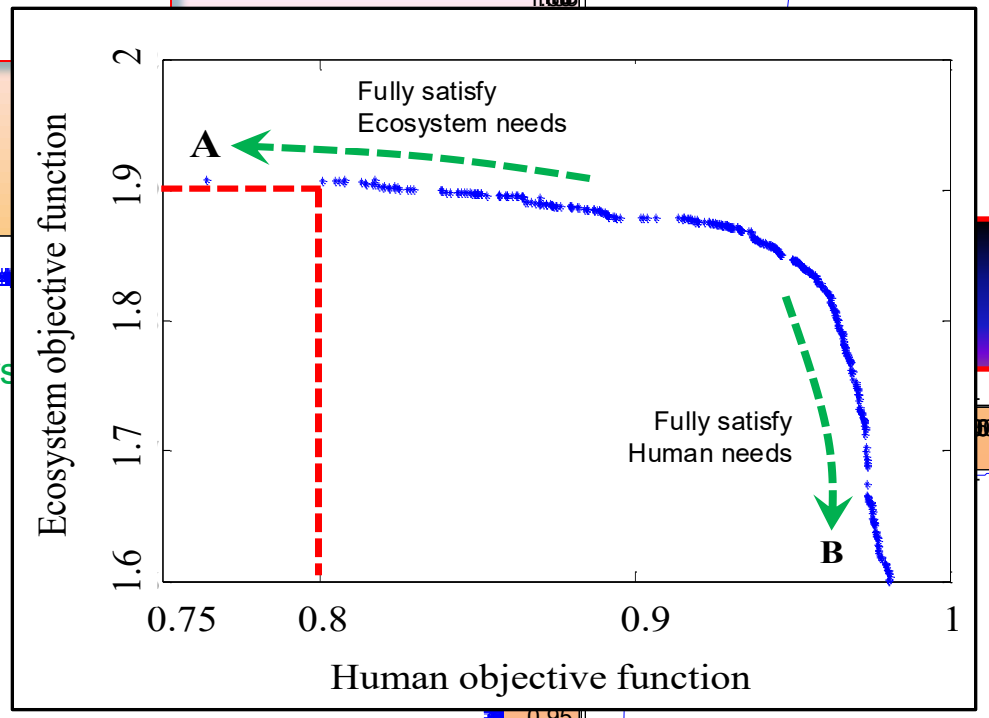
Basin	Flow gauge stations	Elevation (m)	Total species	Number of fish species	Sampling Year	Shannon Index
Danshui River	D1 (U)	249	11	9	1981	1.51
				11	1982	1.53
				7	1983	1.23
	D2 (D)	33	26	18	1981	1.33
				16	1982	1.77
				14	1983	1.55
				16	1997	2.25
				4	1998	2.05
	D3 (D)	30	7	7	1981	1.68
	D4 (D)	1	14	11	1982	1.45
				6	1997	2.02

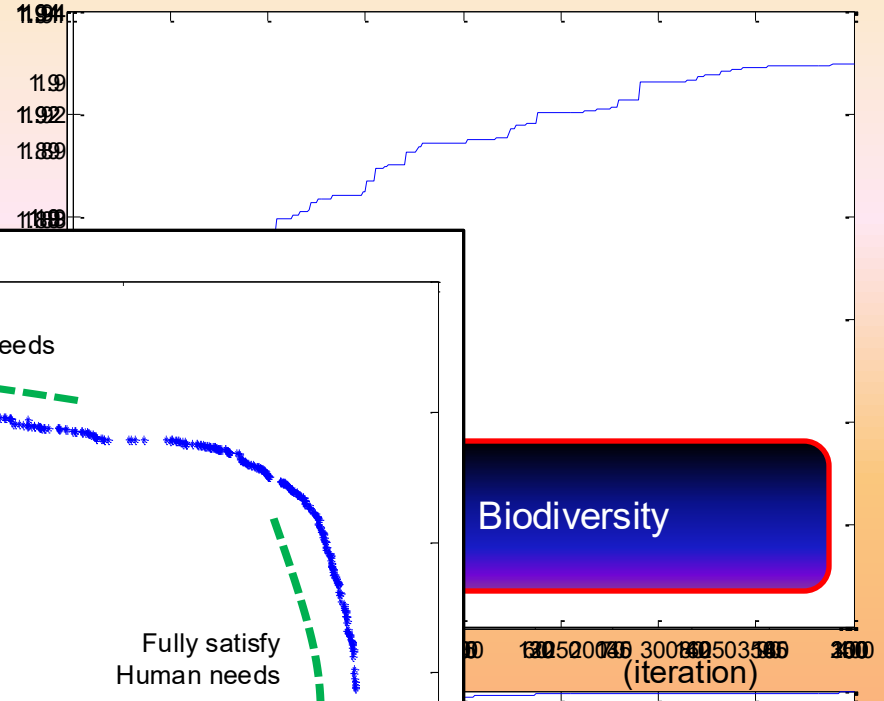
Flowchart of methodology

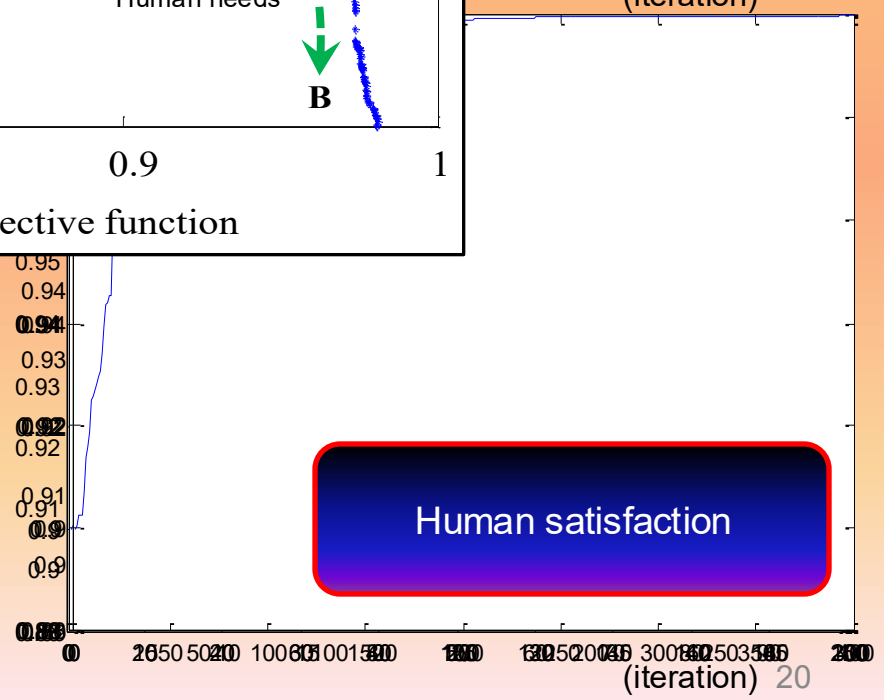


NSGAII with SORBN









PART.02

PENN STATE UNIV.

PennState College of Engineering | CIVIL AND ENVIRONMENTAL ENGINEERING

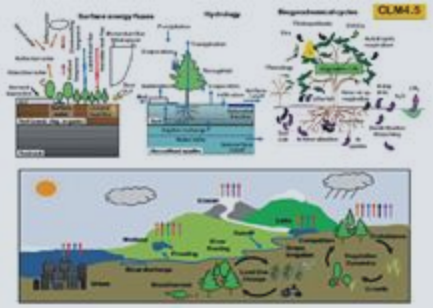
AIMH

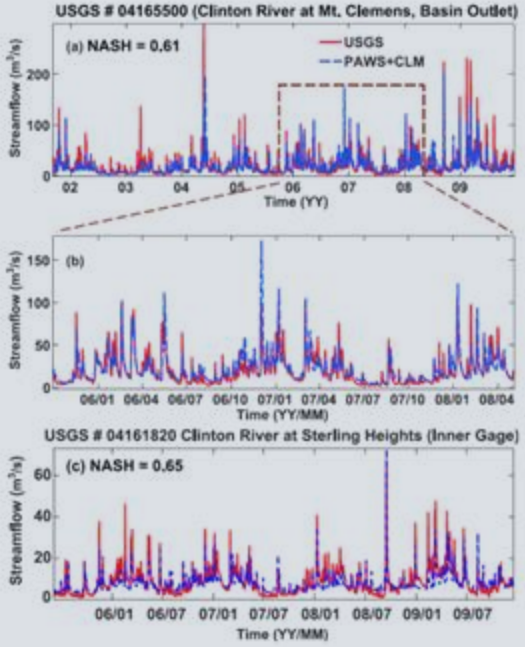
Multi-scale Hydrologic Model

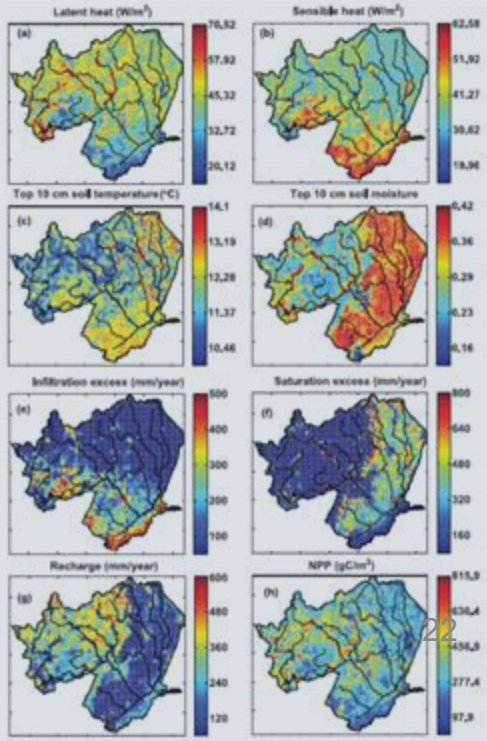
- Process-based Adaptive Watershed Simulator (PAWS) + Community Land Model (CLM)
- Interaction among hydrologic, geographic, and ecosystems.

PAW

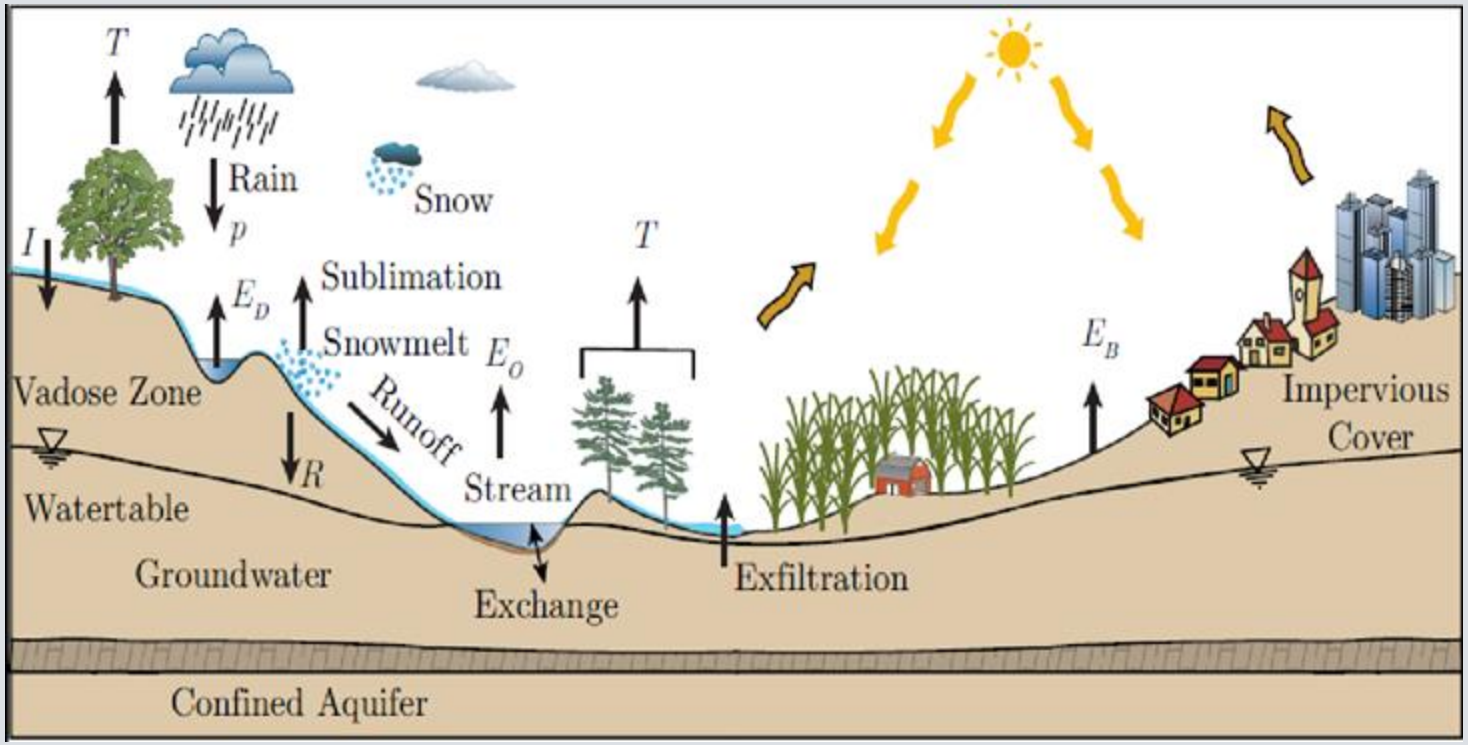
CLM





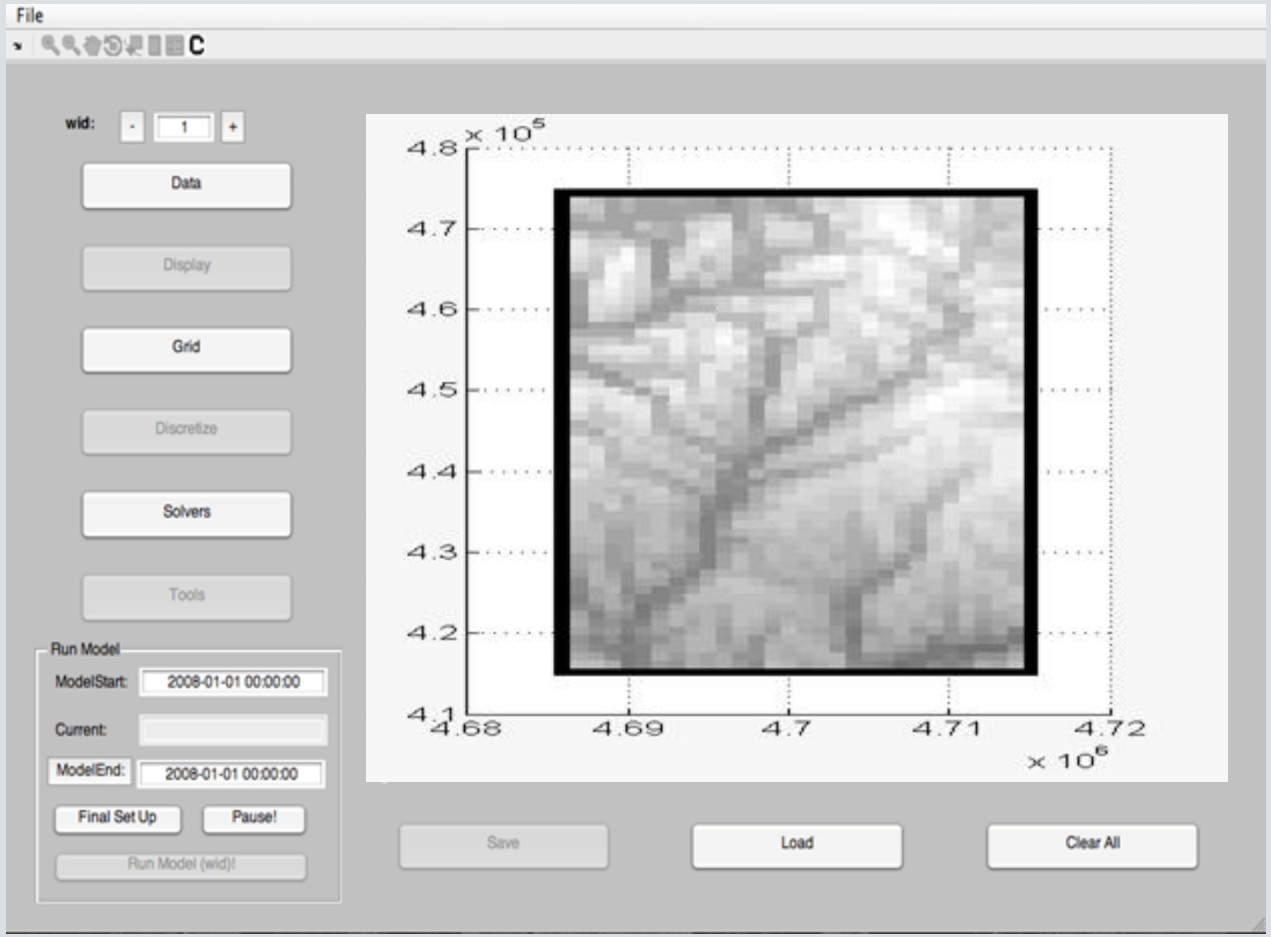


Model domain and processes

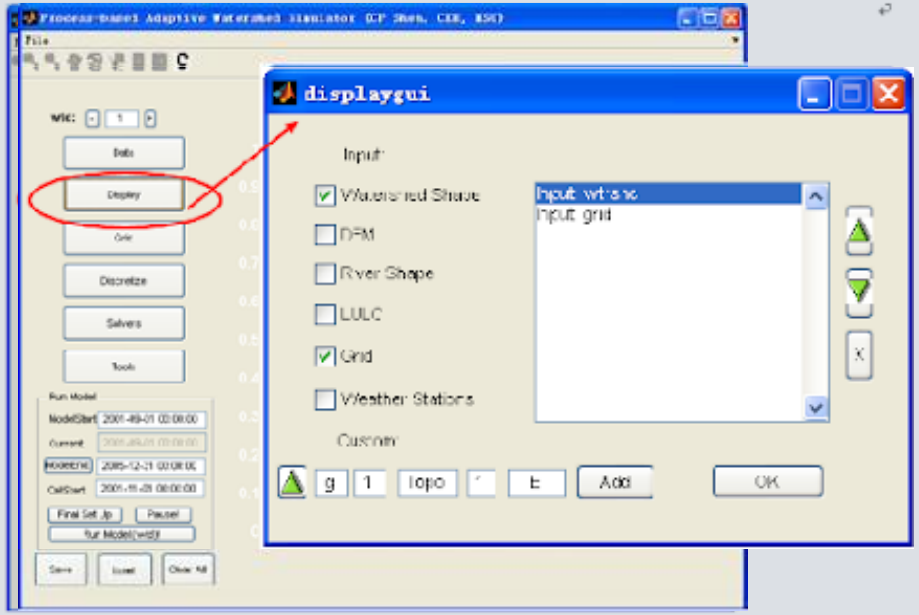


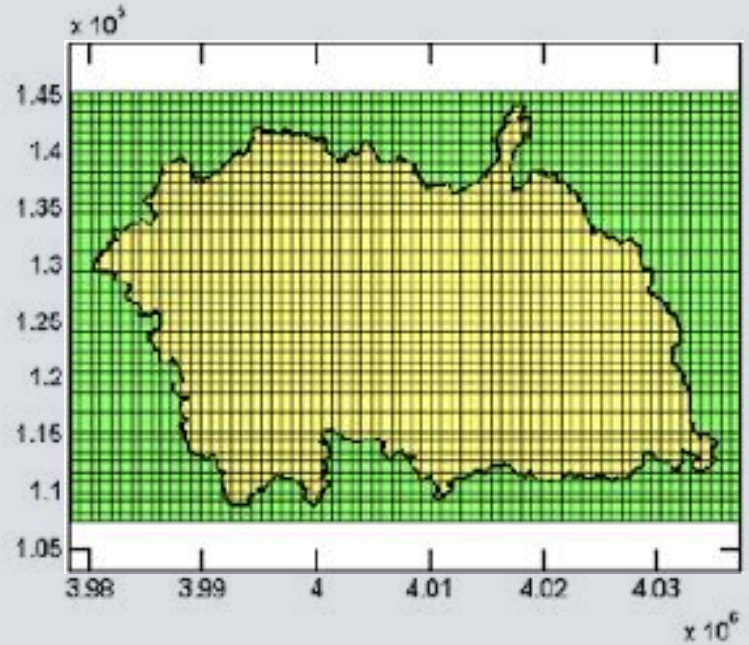
Definition sketch of the model. T: transpiration, p: precipitation, E_O Evaporation from overland flow/stream, E_B: evaporation from bare soil; I: infiltration; R: Recharge

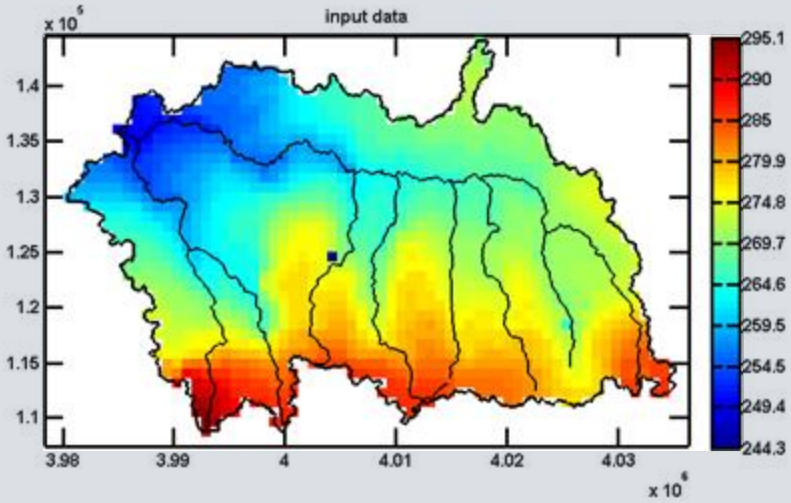
GUI interface

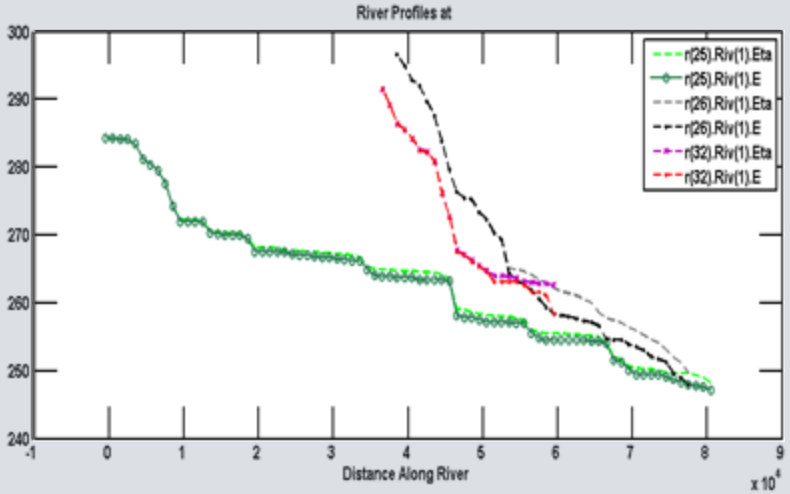


PENN STATE UNIV.

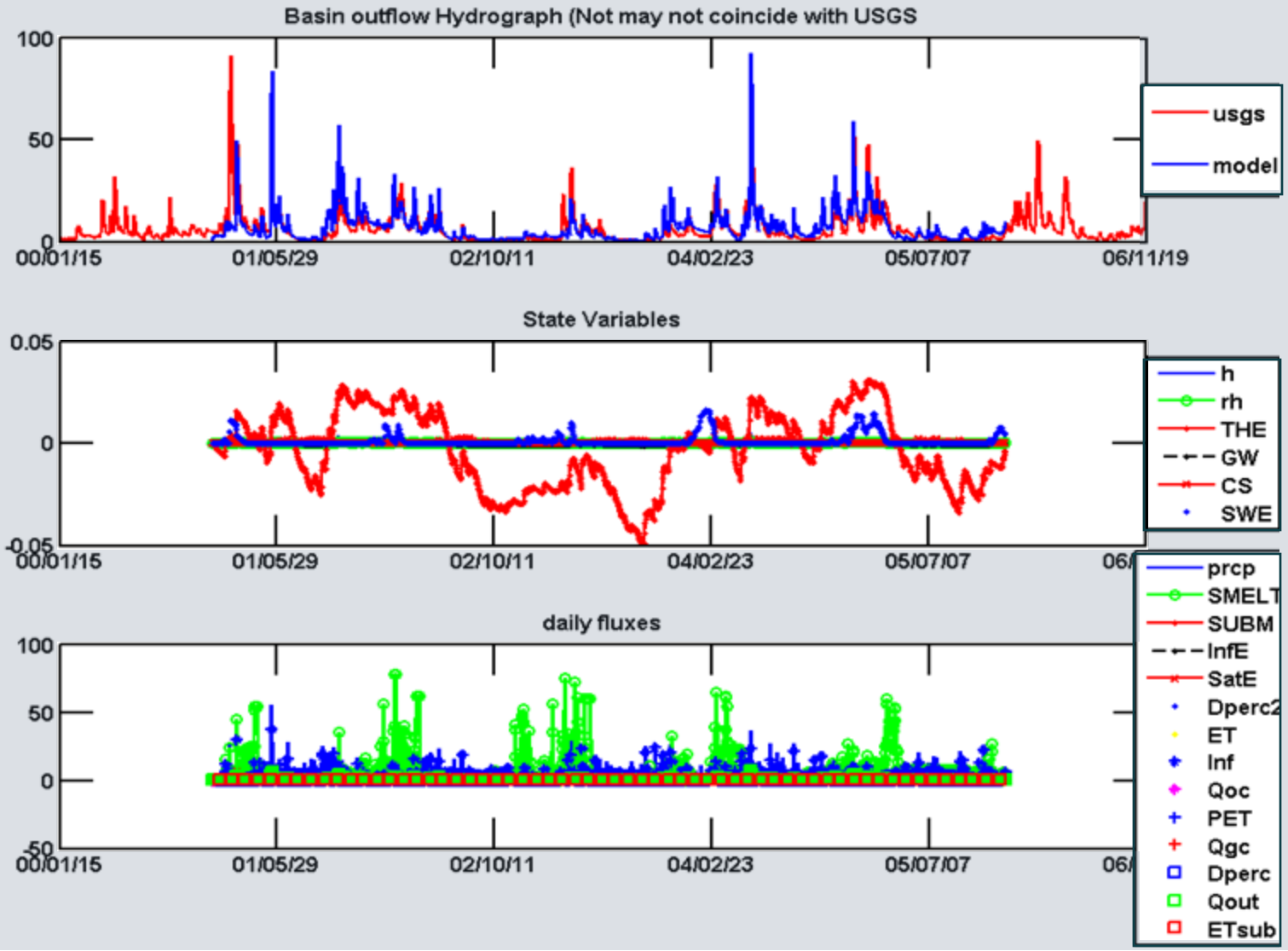






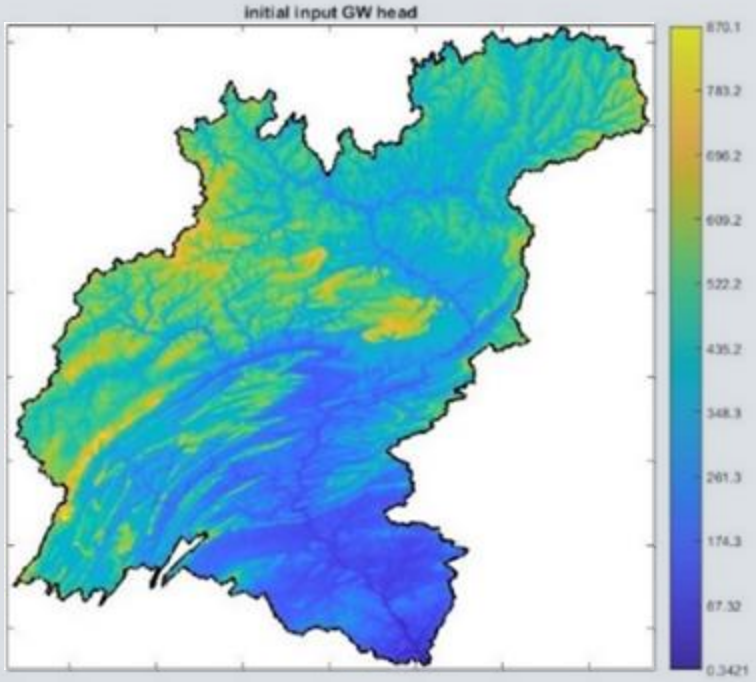


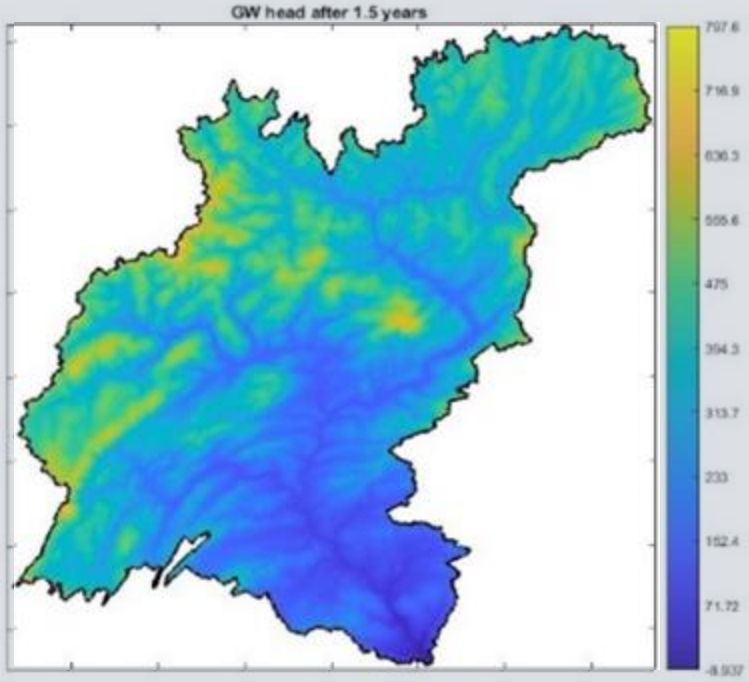
PENN STATE UNIV.



Susquehanna River Basin (SRB)

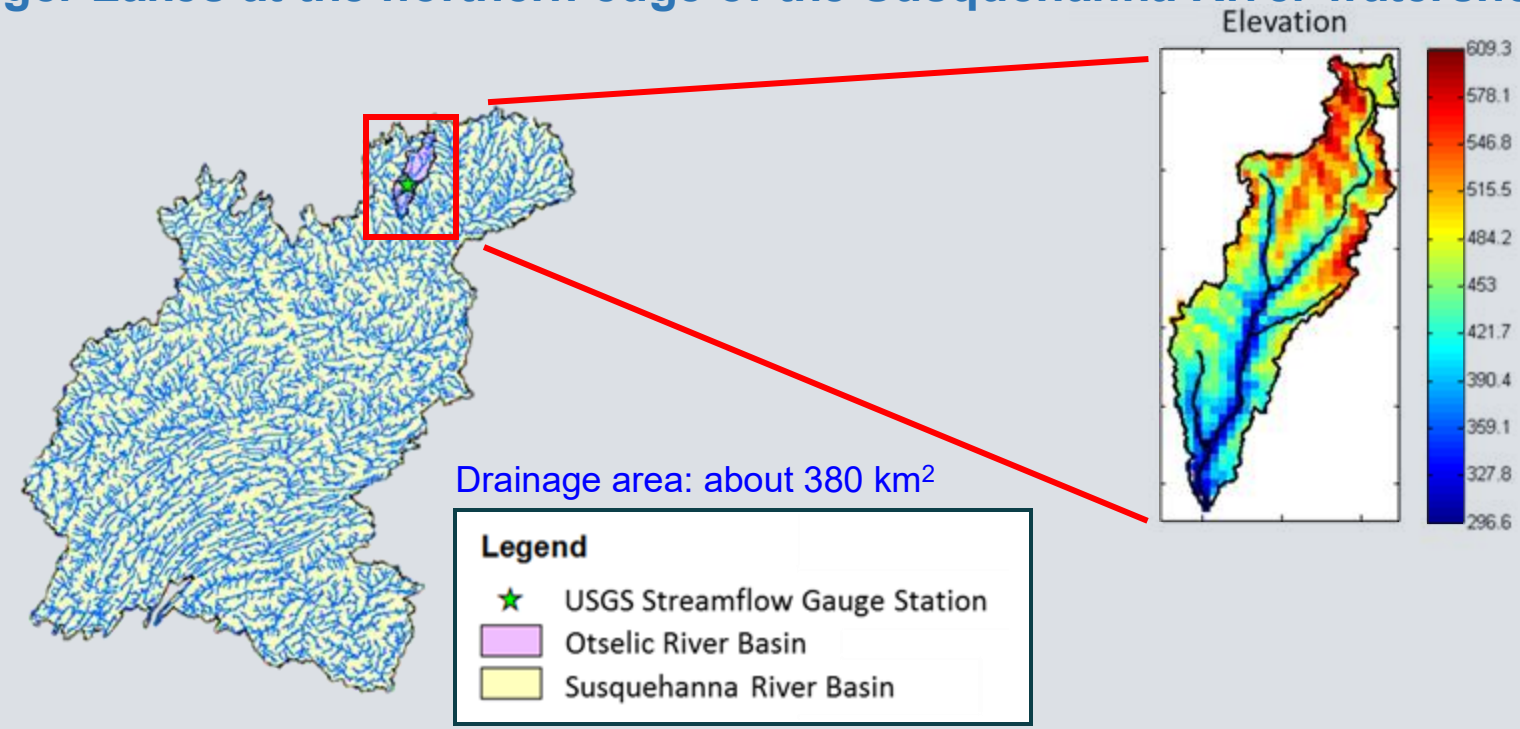
Drainage area: about 72,000 km^2





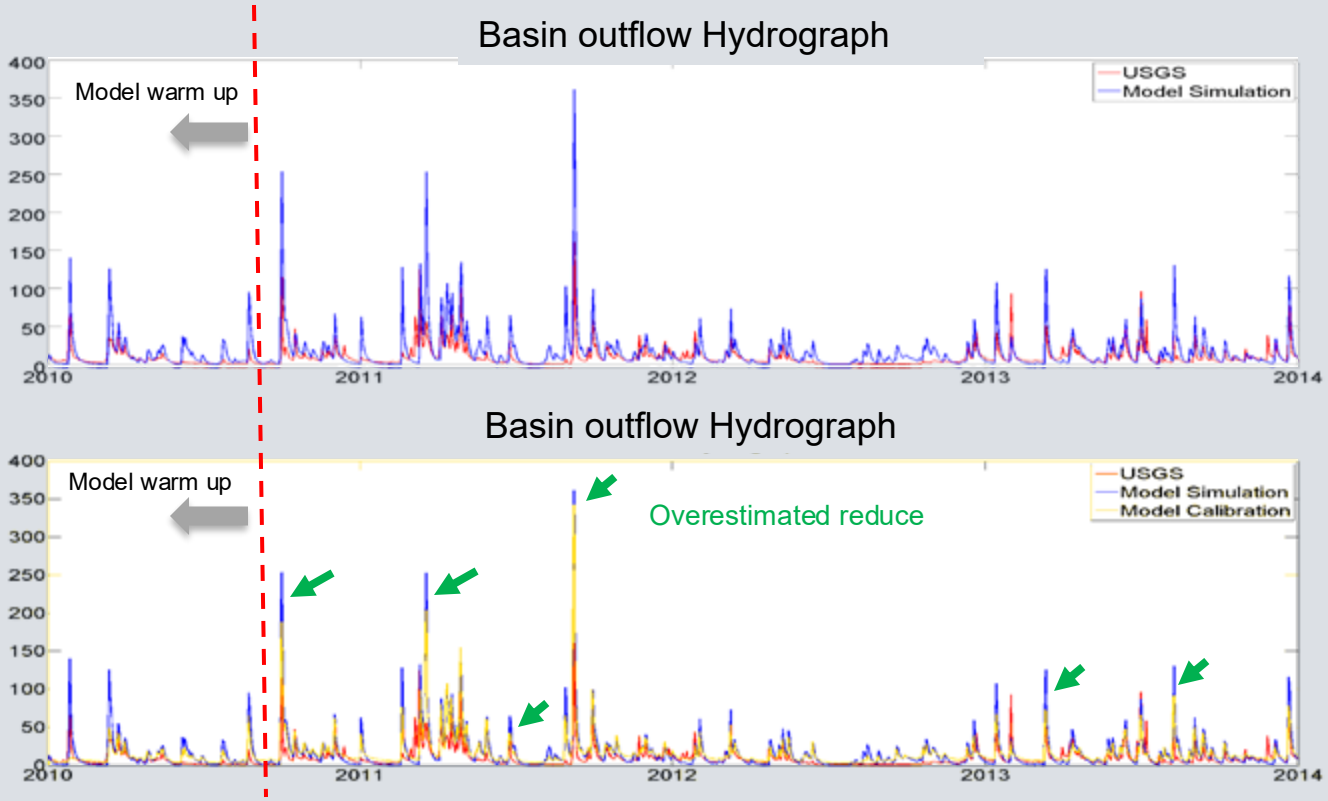
Otselic River Basin

- The Otselic River drains a mountainous area, mostly forested and agricultural, east of the Finger Lakes at the northern edge of the Susquehanna River watershed



Results of Simulation and calibration

Forcing data: 2010-2013 (4 years)

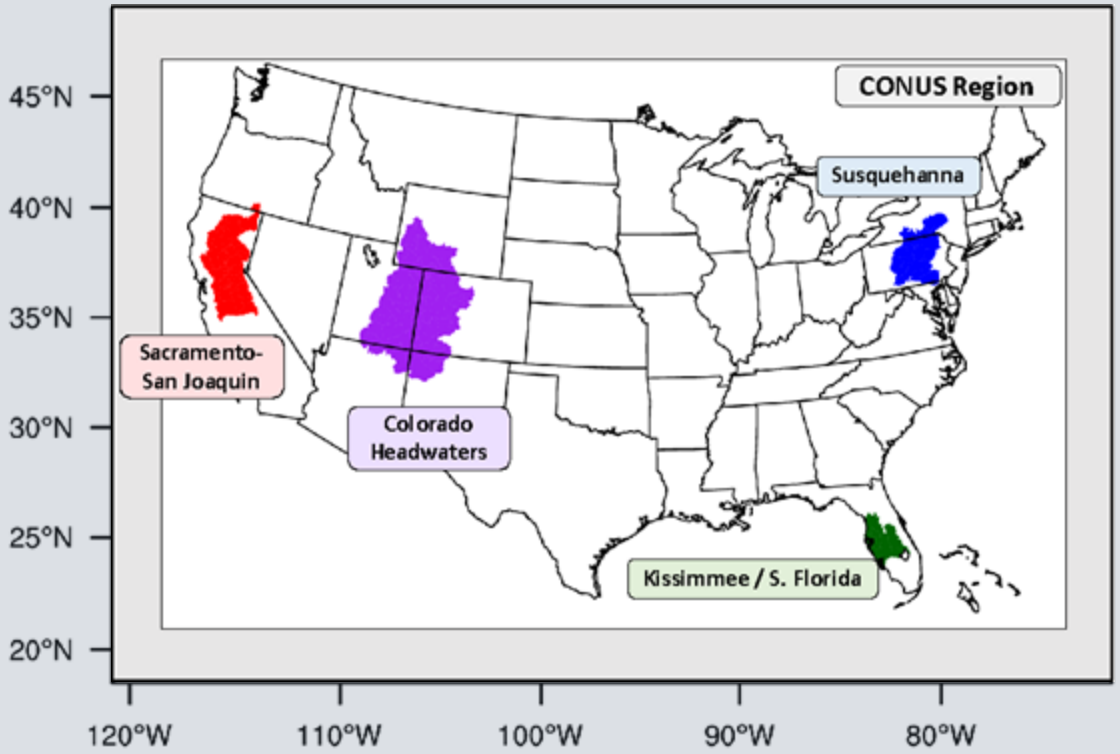


02 PENN STATE UNIV.

HYPERION

UNDERSTANDING HYDROCLIMATE DATA WITH USE-INSPIRED METRICS

Stony Brook University | University of Delaware | UtahState University | UCLA | PennState | cri Climate Readiness Institute | Pacific Northwest National Laboratory | UC Davis University of California | Berkeley Lab | NCAR



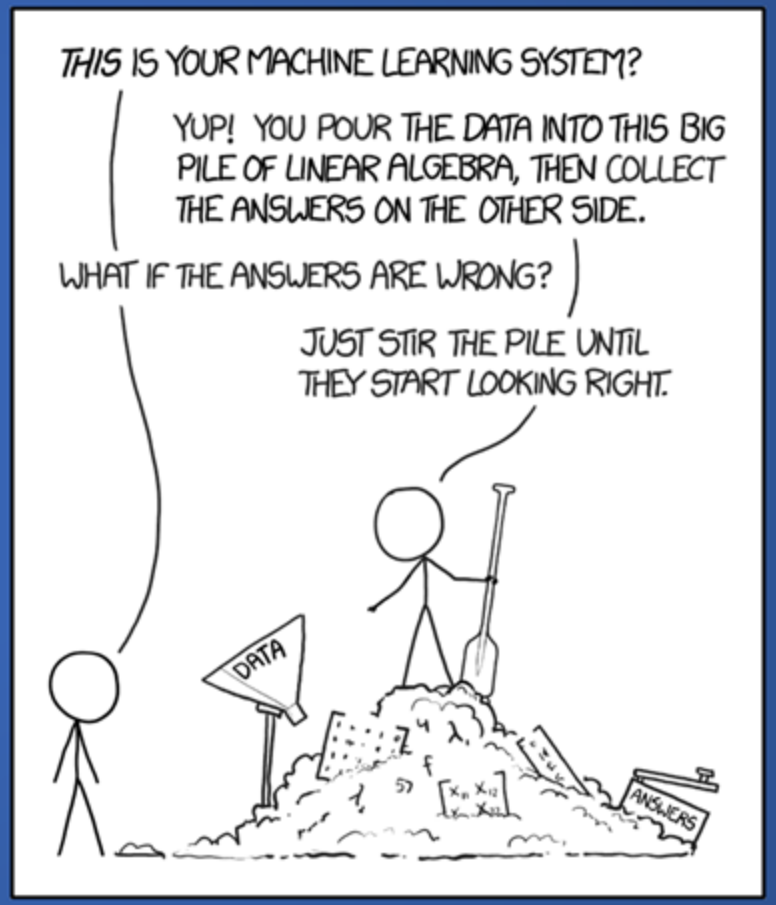
PART.03

Integrating process-based models and machine learning

Integrating process-based models and machine learning

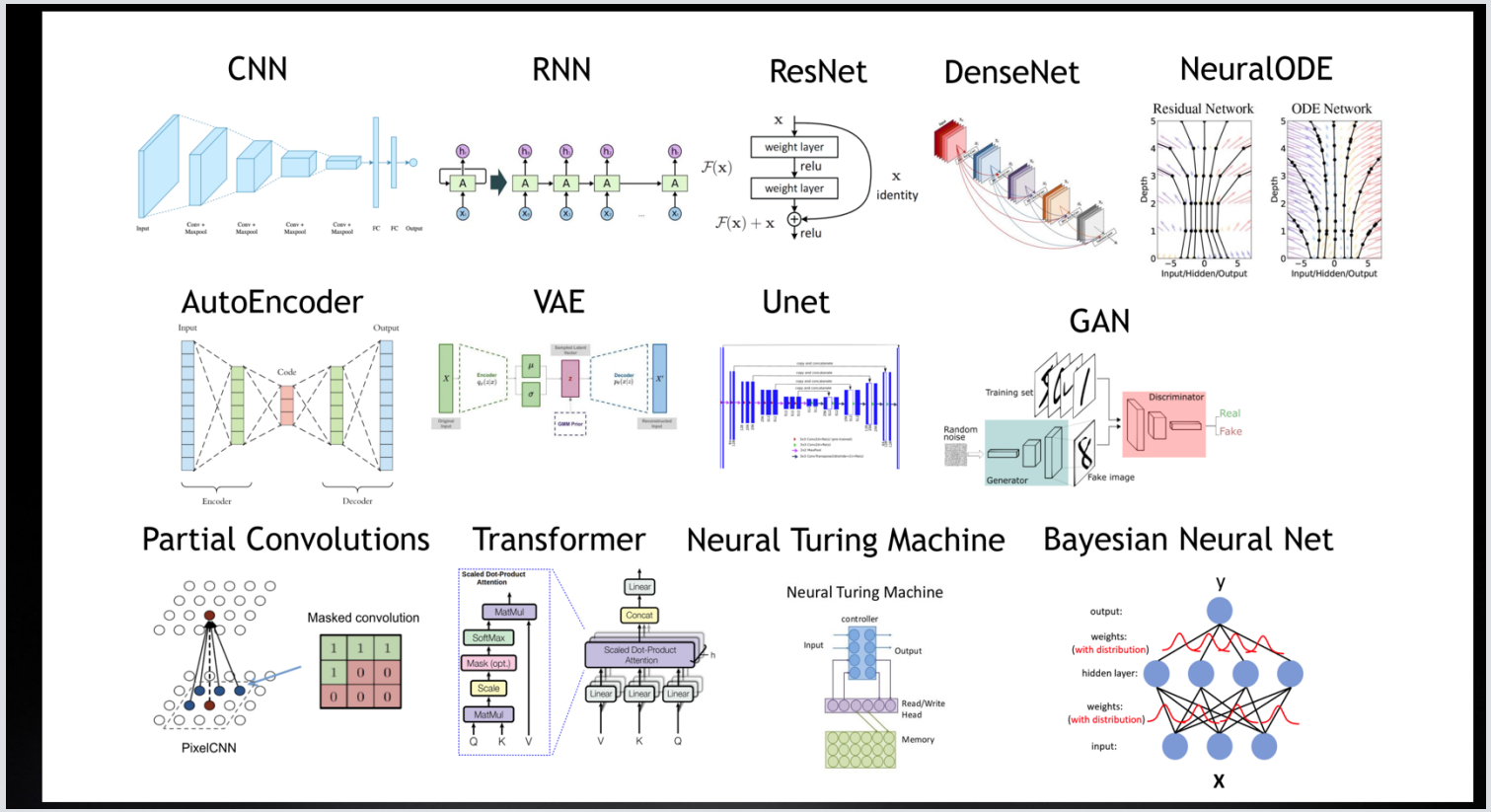
- ML and deep learning have become popular
  - Not always the right tool!
- Misconception is that ML is a "black boxes"
- **Goal: demystify ML and deep learning models by demonstrating benefits and drawbacks of model interpretation and visualization (MIV)**

What the public thinks ML is:

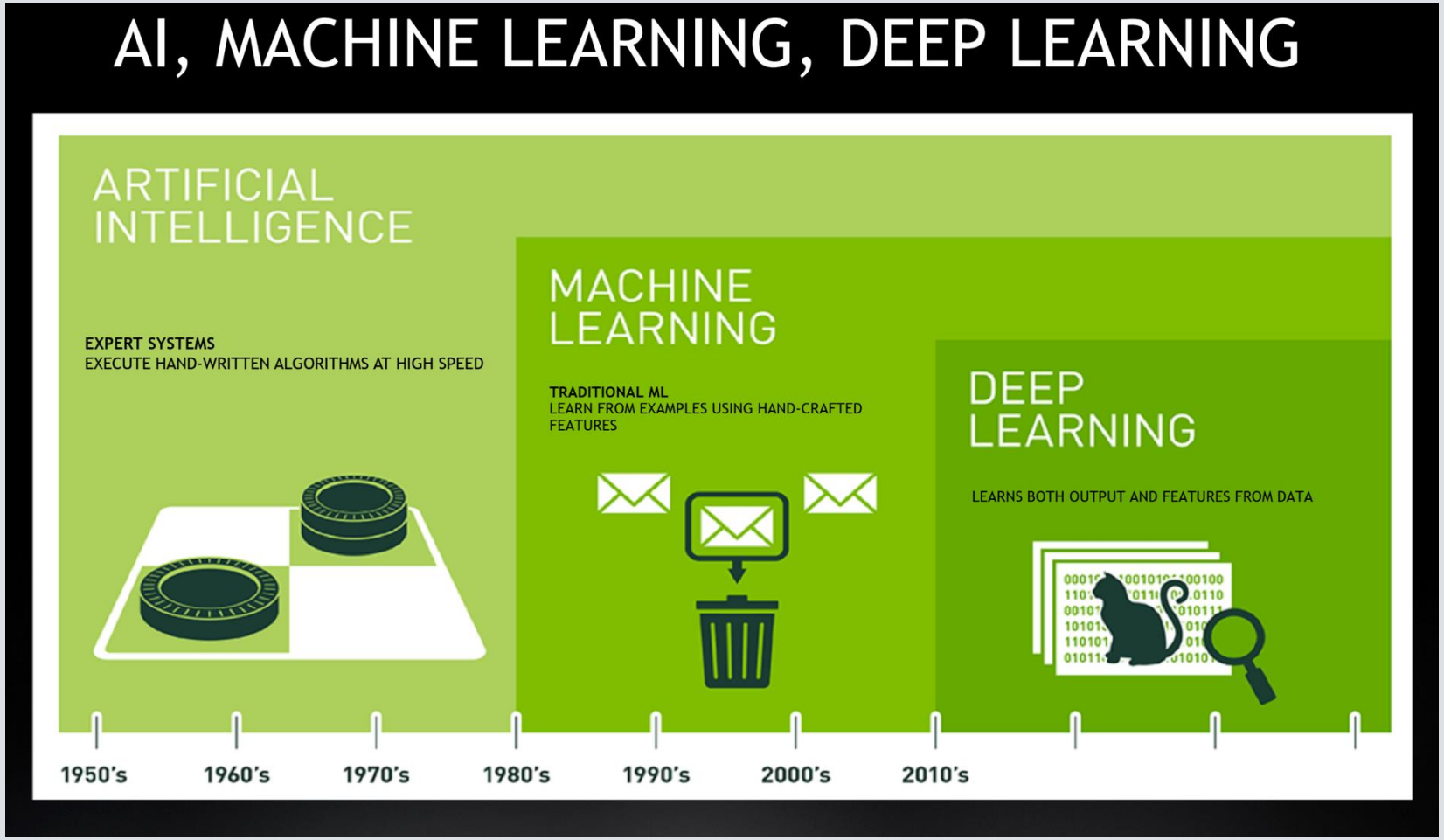


https://xkcd.com/1838/

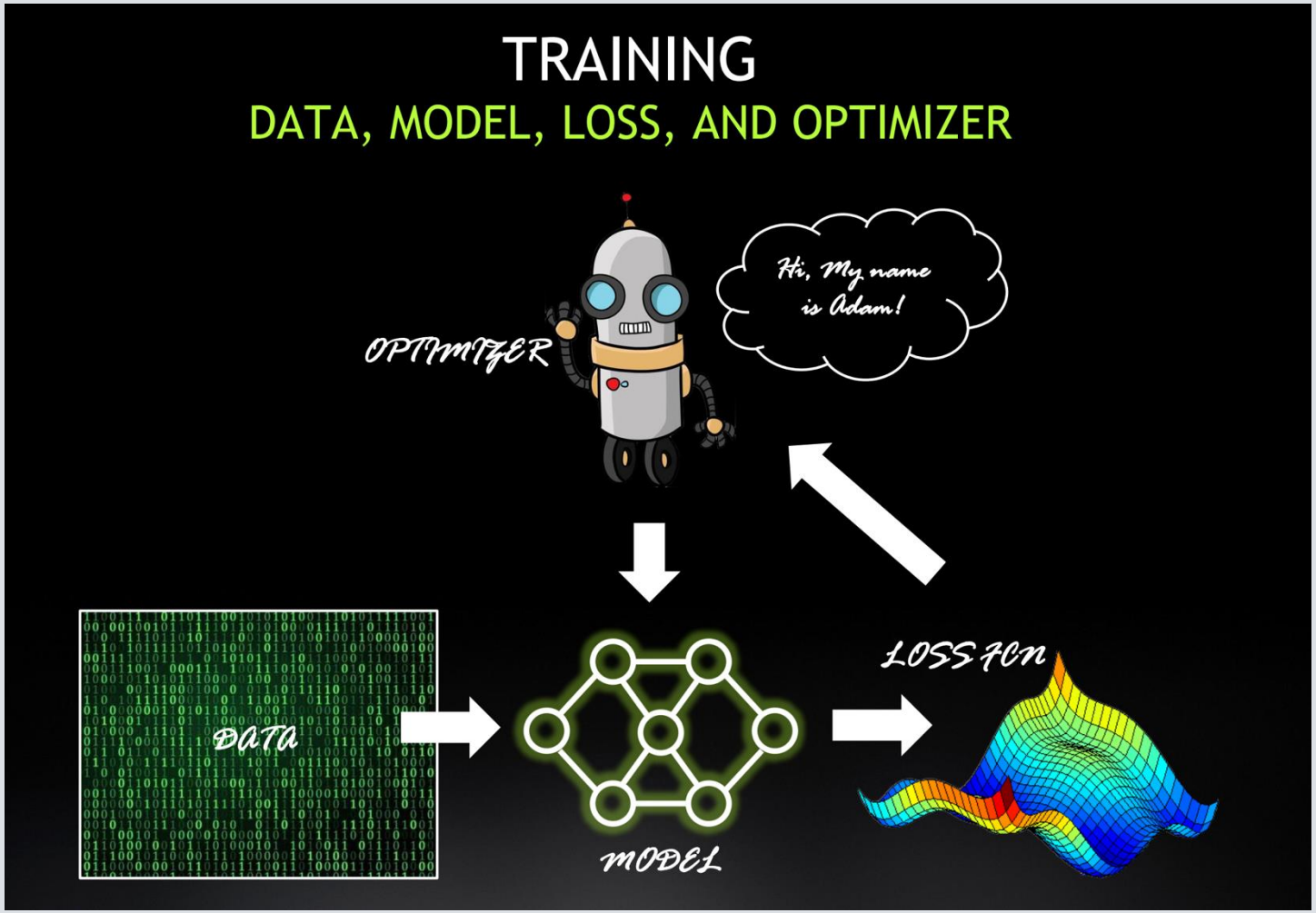
Machine learning (ML)/Deep Learning (DL)



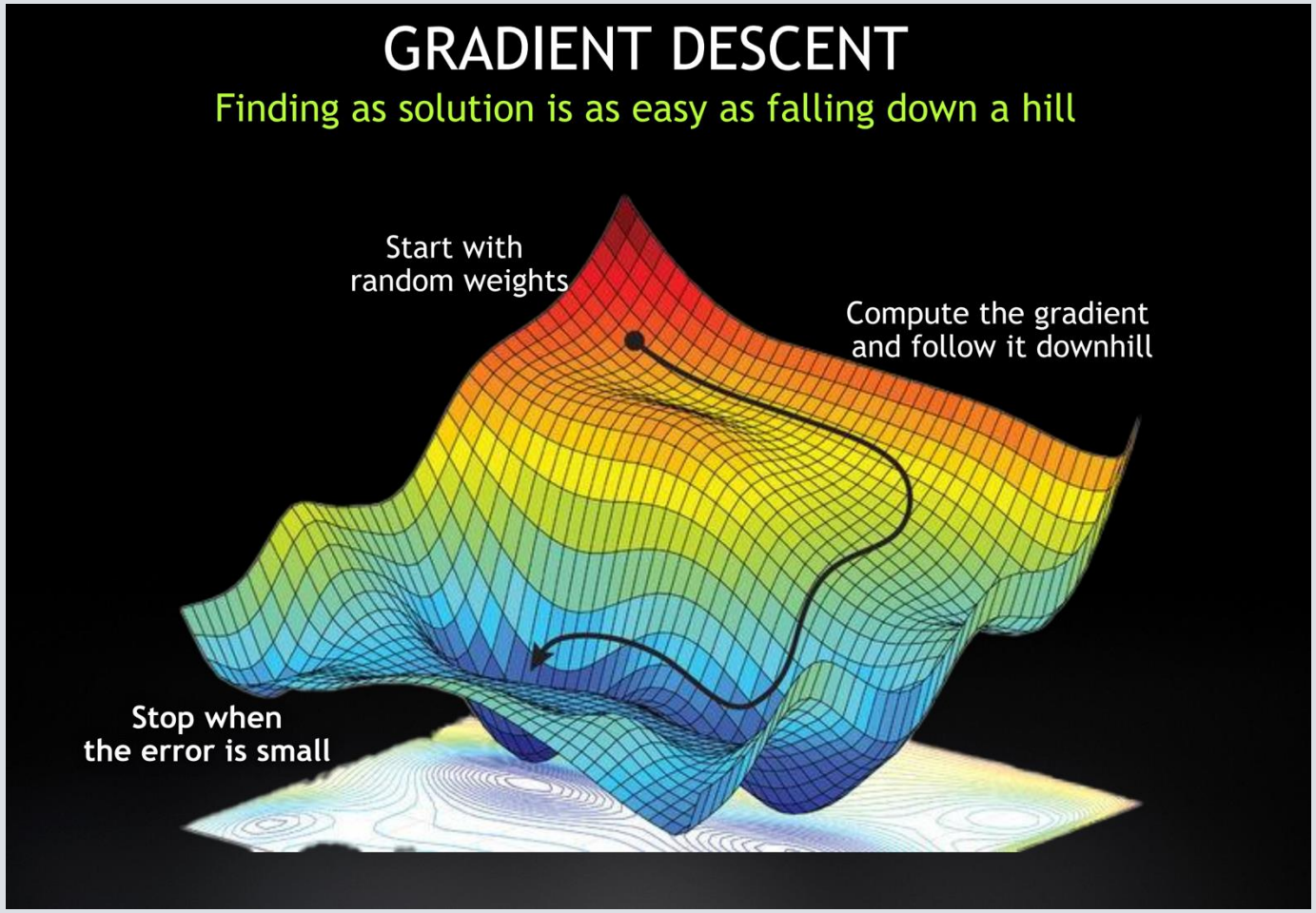
Integrating process-based models and machine learning



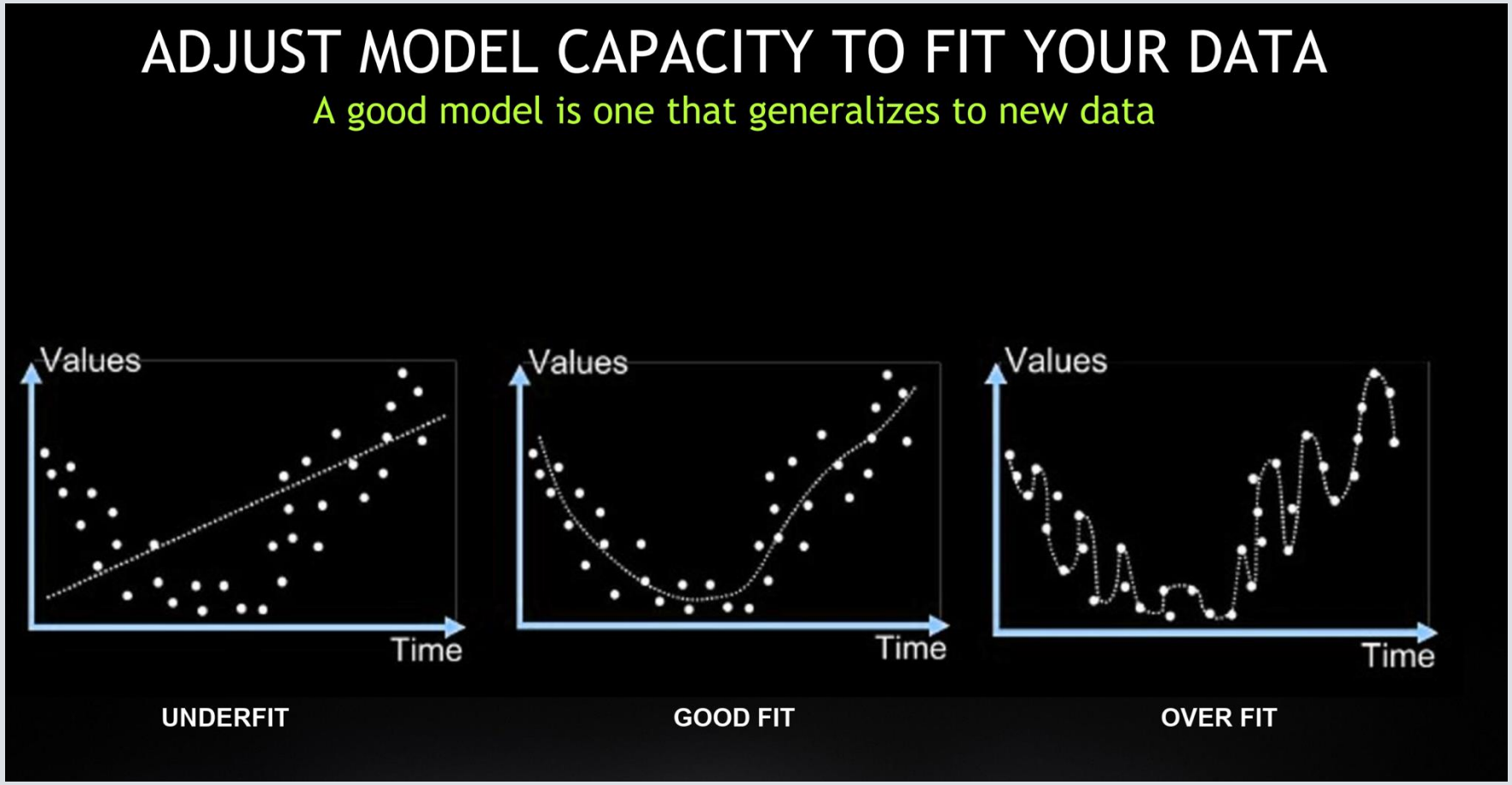
Integrating process-based models and machine learning



Integrating process-based models and machine learning



Integrating process-based models and machine learning



Projects

Deep Learning -- Success stories

Dexterity, OpenAI, 2019

GANs Face Generation, becominghuman.ai, 2019

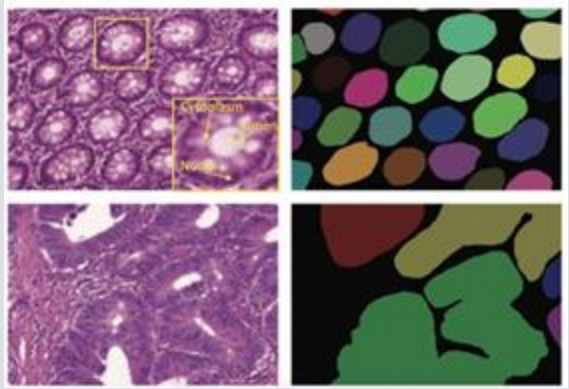
https://openai.com/blog/learning-dexterity/

Deep Learning -- Success stories

Self-driving Cars

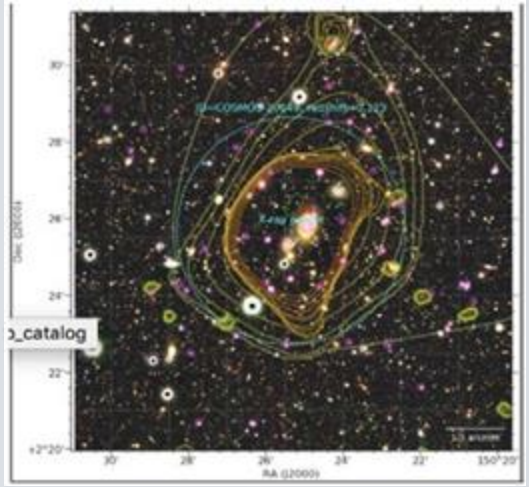
AI art and music

Deep Learning -- Success stories in science



The top row shows benign tissue that has been segmented so it is easier to analyze. The bottom row shows abnormal tissue. [Courtesy of the Chinese University of Hong Kong.]

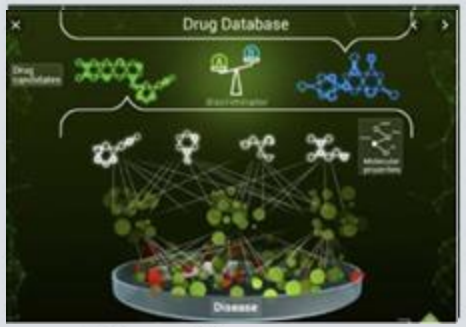
Cancer detection



Mapping the universe

Predict protein structure

Land cover segmentation



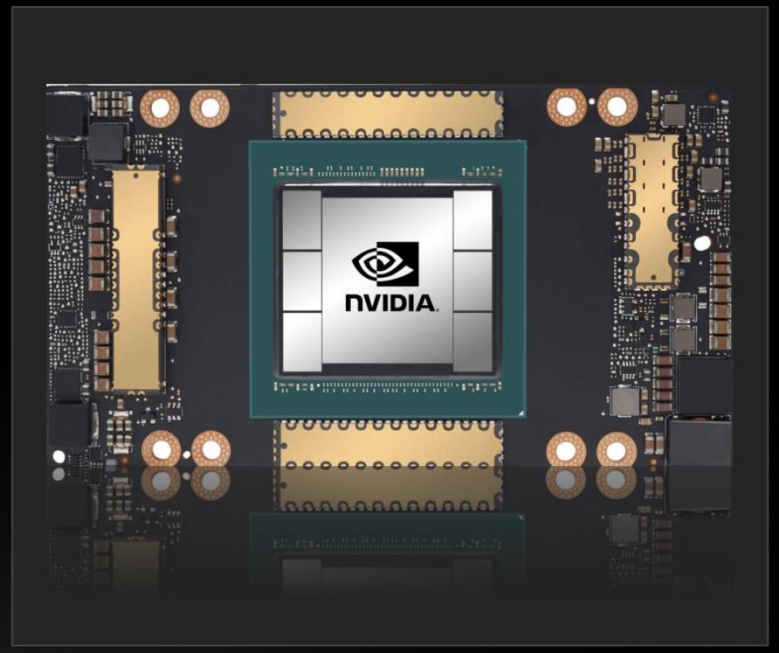
Drug discovery

WHAT YOU NEED TO MAKE IT WORK

You need three main ingredients (and some skill)

LARGE QUANTITIES OF DATA

ML FRAMEWORK



GPU ACCELERATOR

DEEP LEARNING FRAMEWORK

Many frameworks to choose from (but not Fortran)

PyTorch

K + TensorFlow

Python

mxnet

GLUON

C++

flux

Julia

Chainer | mxnet | Caffe2 | Microsoft CNTK | TensorFlow | K Keras | GLUON | PYTORCH | theano

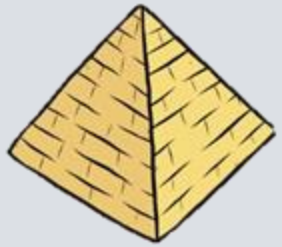
Process-Based Modeling (PBM) vs Machine Learning (ML)

===== PBM strength =====

- Build from the bottom-up to observe emerged patterns
- We know what we put in
- We can do experiments & identify causal relationships

===== Limitations ====

- Human biases
- Parameters calibration
- What if the process we don't know?
- Errors compound?



Bottom-up

Synergy

===== Machine learning strength =====

- Build from the top-down, directly from observations
- Less Human biases
- Help identify things we don't know
- Highly efficient in computation

==== Limitations ======

- May be difficult to interpret
- Can't observe everything!
- May not fully respectphysical laws
- Does not understand causal relationships!

Top-down

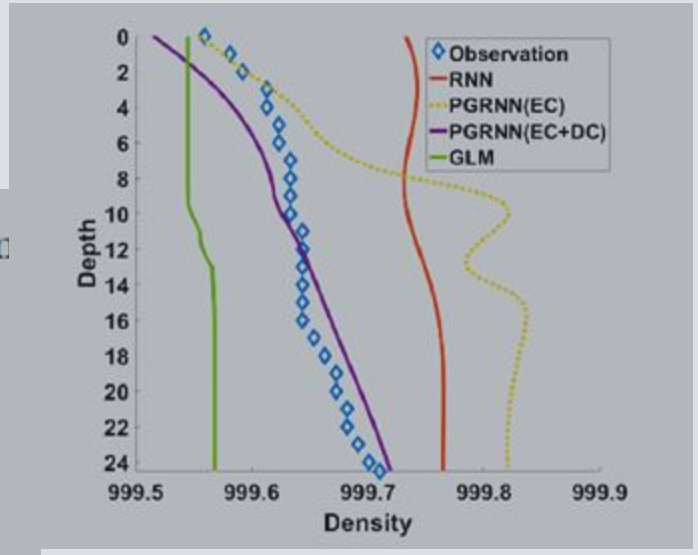
Physics(theory)-guided machine learning

- Encoding laws into the loss function

$$\mathcal{L} = \boxed{\mathcal{L}_{\mathrm{RNN}}} + \lambda_{\mathrm{EC}}\mathcal{L}_{\mathrm{EC}} + \lambda_{\mathrm{DC}}\mathcal{L}_{\mathrm{DC}}$$

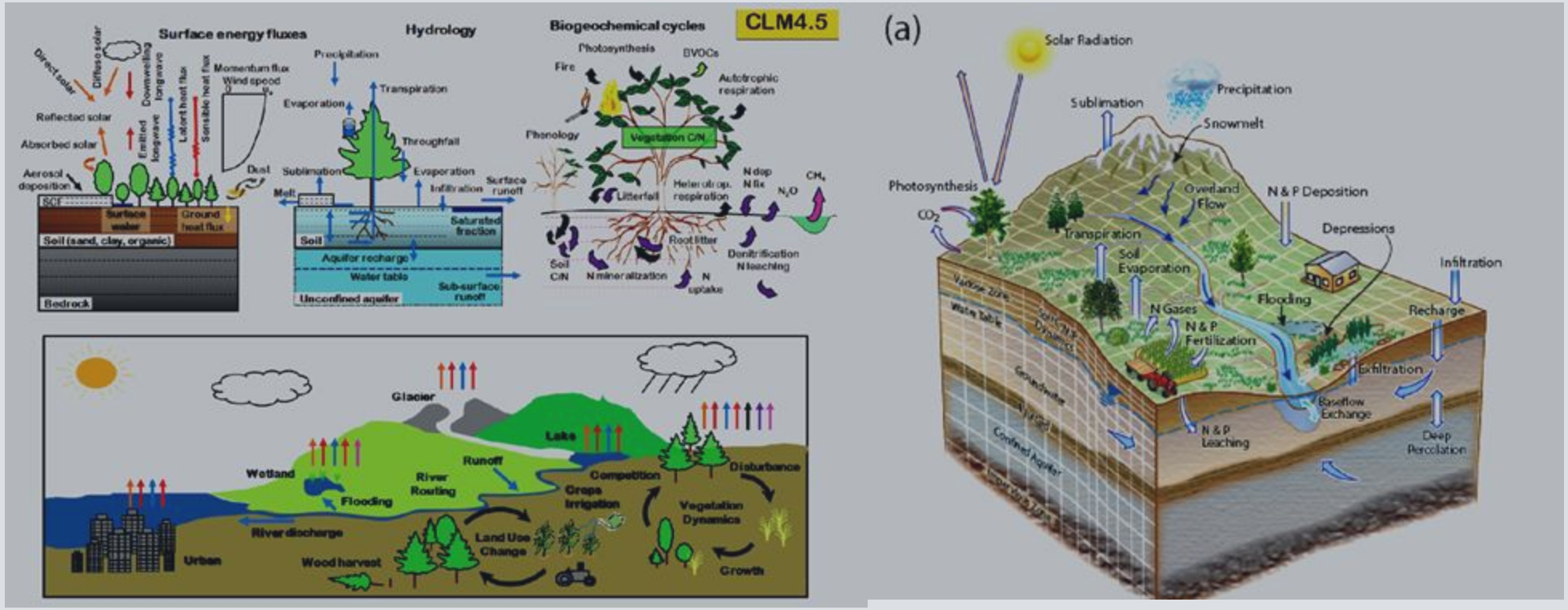
Physics Guided RNNs for Modeling Dynamical Systems: A Case Study in Simulating Lake Temperature Profiles

Xiaowei Jia[1], Jared Willard[1*], Anuj Karpatne[2], Jordan Read[3], Jacob Zwart[3], Michael Steinbach[1], Vipin Kumar[1]
[1]Department of Computer Science and Engineering, University of Minnesota
[2]Department of Computer Science and Engineering, Virginia Tech
[3]U.S. Geological Survey
[1]{jiaxx221,willa099,stei0062,kumar001}@umn.edu, [2]karpatne@vt.edu, [3]{jread,jzwart}@usgs.gov

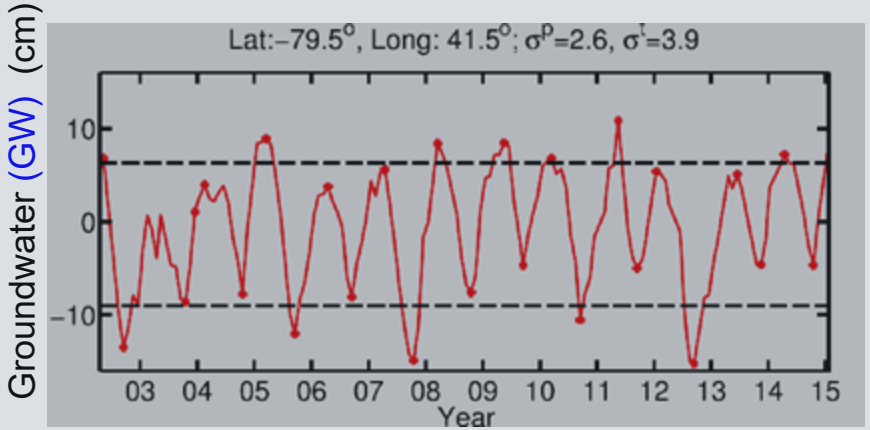


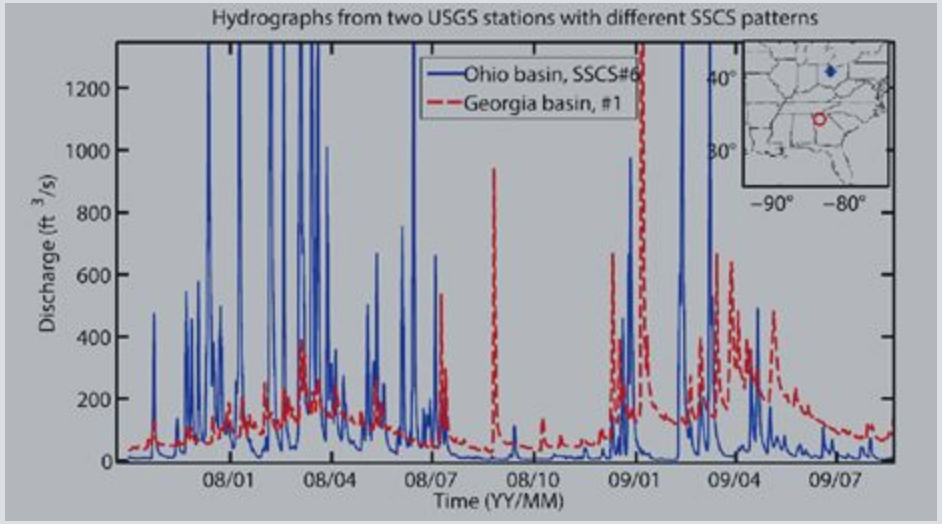
Tend to generalize better

But what if the model was too complicated?



The storage streamflow correlation spectrum (SSCS) as a revealing hydrologic signature





Calculated R(*GW*, Q_x) for different percentiles x

Water Resources Research

RESEARCH ARTICLE

10.1002/2016WR020283

Key Points:

- Utilizable full-flow-regime streamflow-storage correlations exist under diverse climates, but the patterns vary with environmental

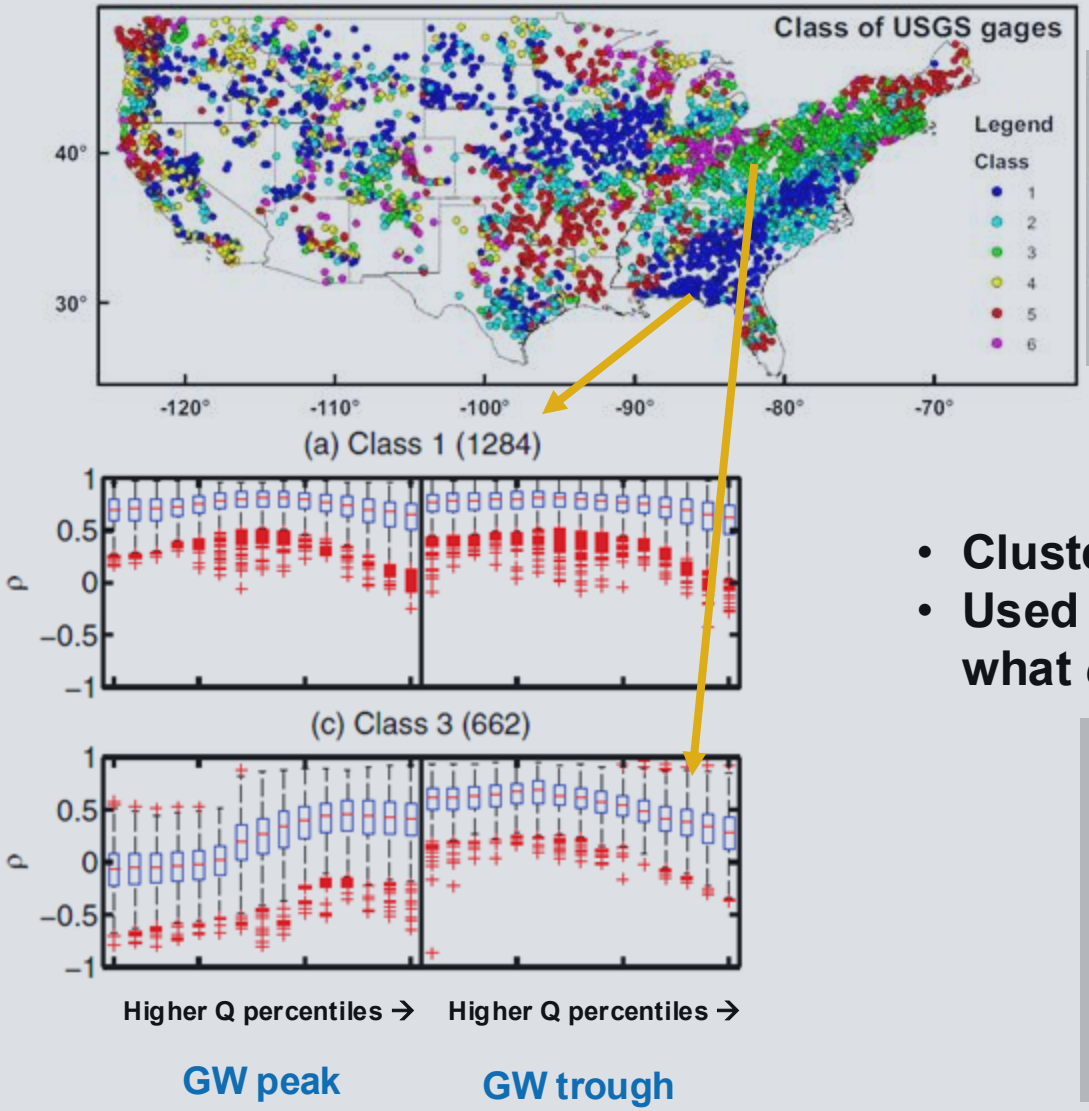
Full-flow-regime storage-streamflow correlation patterns provide insights into hydrologic functioning over the continental US

Kuai Fang[1] and Chaopeng Shen[1]

[1]Civil and Environmental Engineering, Pennsylvania State University, University Park, Pennsylvania, USA

Fang and Shen, 2017, WRR

Integrating process-based models and machine learning



Water Resources Research

RESEARCH ARTICLE

10.1002/2016WR020283

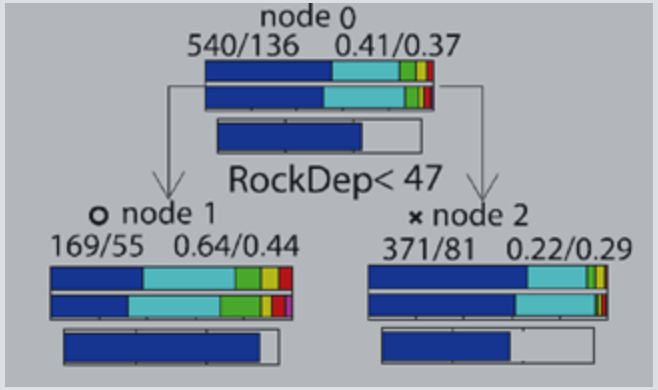
Full-flow-regime storage-streamflow correlation patterns provide insights into hydrologic functioning over the continental US

Key Points:
- Utilizable full-flow-regime streamflow-storage correlations exist under diverse climates, but the patterns vary with environmental

Kuai Fang[1] and Chaopeng Shen[1]

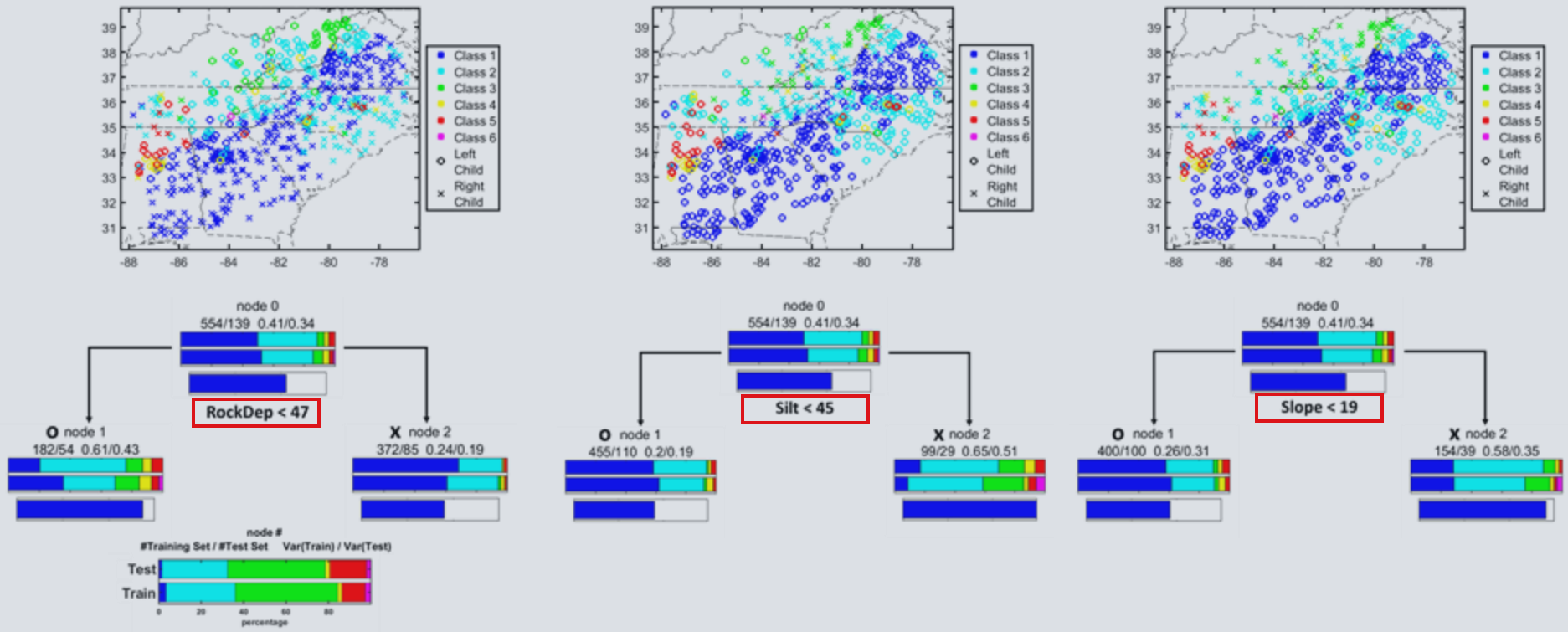
[1]Civil and Environmental Engineering, Pennsylvania State University, University Park, Pennsylvania, USA

- Clustered them into 6 classes
- Used Classification tree to understand what *caused* their differences.

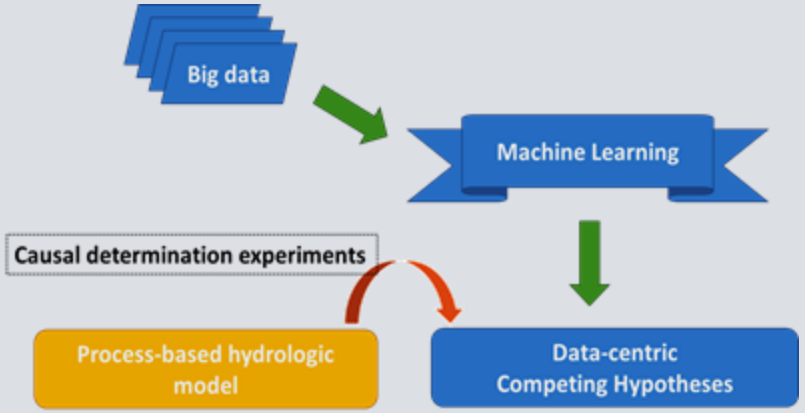


What data mining doesn’t know……

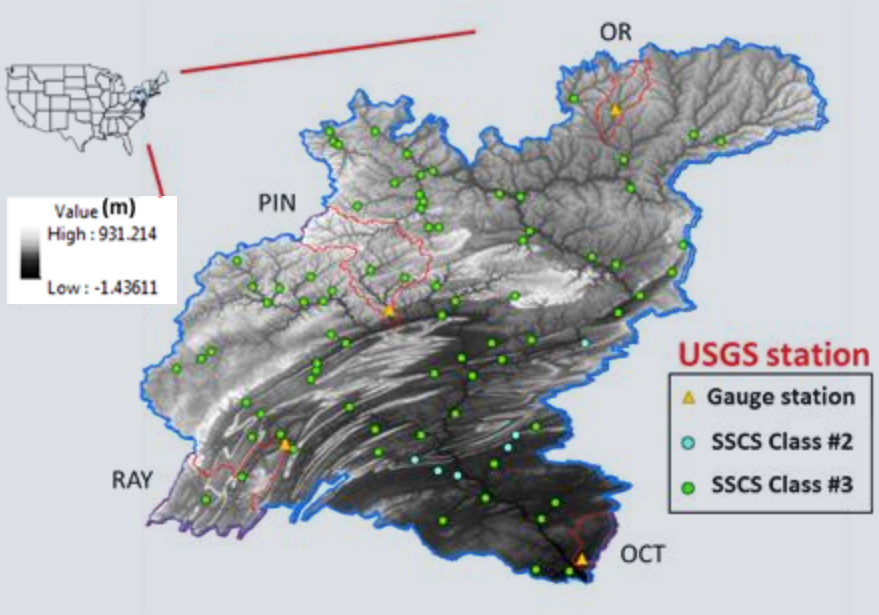
- **Factors covary……Hard to distinguish between three competing hypotheses!**

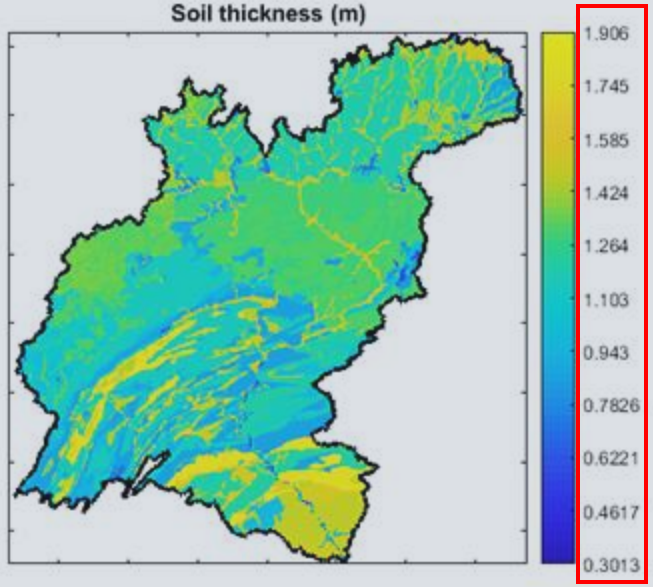


Experimental area –Susquehanna River Basin (SRB)



SRB is a major river located in the northeastern United States (watershed area: 71,225 km^2)



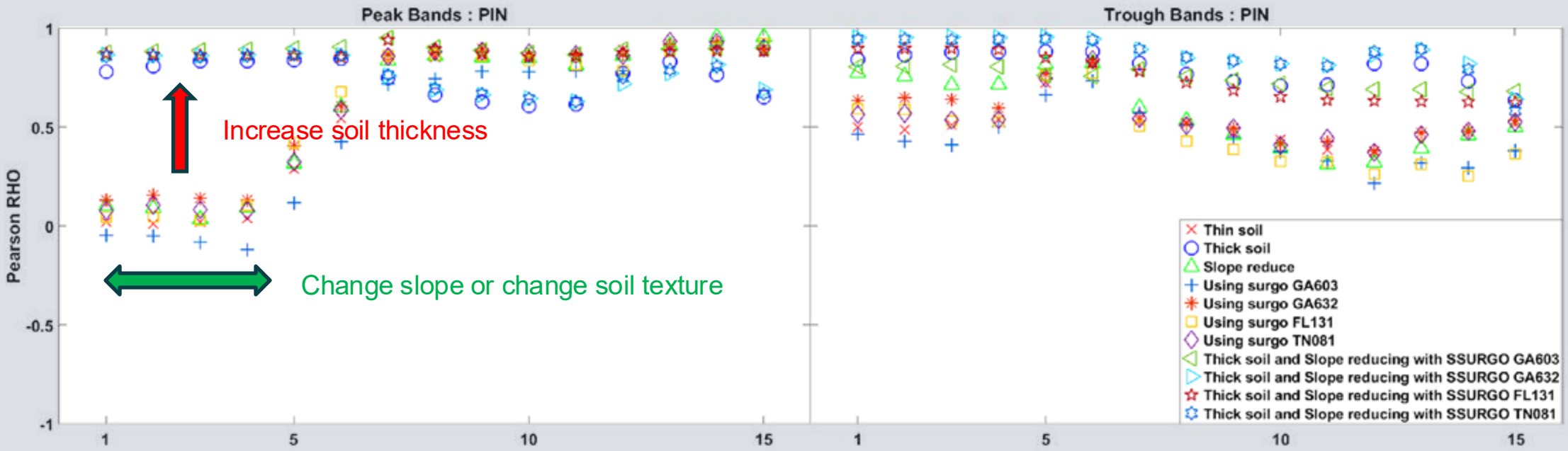


Thin soil thickness

We run perturbed experiments with increased soil thickness & other variables

Experiment results of the storage streamflow correlation

We tested: (1) Increase soil thickness; (2) Swap soil types; & (3) change terrain slope



Higher Q percentiles →

GW peak

Higher Q percentiles →

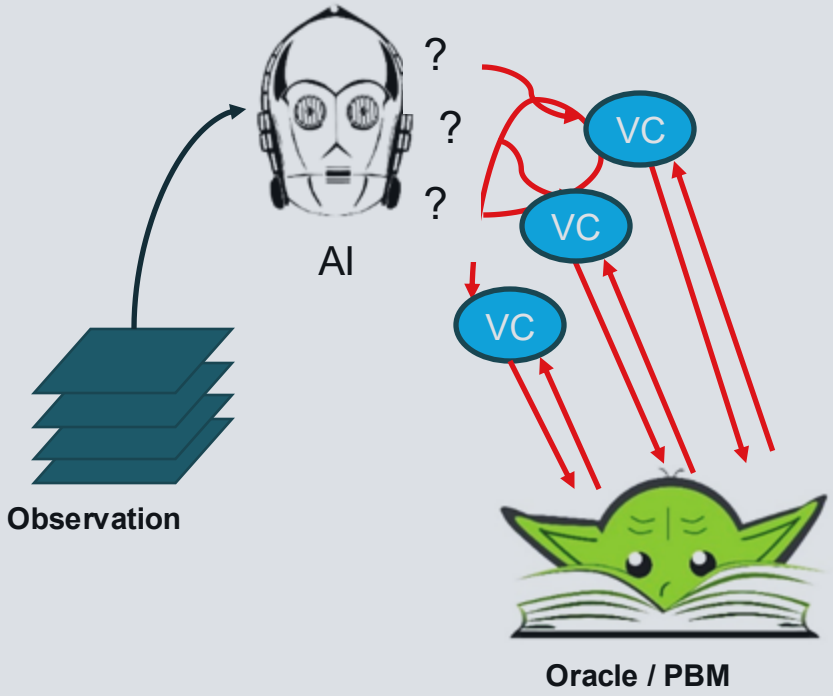
GW trough

What is cool about this this data-centric framework?

It can be completely automated!

What we need to do:

1. Define an AI agent that generate hypotheses from data
2. Let this agent have access to a PBM as an “Oracle” (set up and run experiment, and retrieve results)
3. Define the prior & the likelihood function
4. Let the robot go through all the data and deliver us knowledge!



Tsai et al., 2020, Revealing causal controls of storage-streamflow relationships with a data-centric Bayesian framework combining machine learning and process-based modeling. *Frontiers in Water. 2, 40.*

Summary

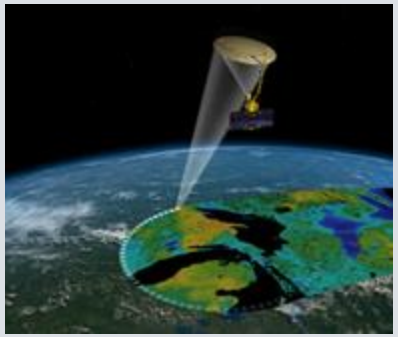
- The hypothesis that large soil thickness causes higher storage-streamflow correlation is true, while other factors are only coincidental factors
- The data-centric framework is one of the possible options to integrate data and complex PBMs
- This framework is highly efficient and more objective

PART.04

From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling

AIMH

Background



- **Soil Moisture Active Passive (SMAP)**
  - Launched from 2015/04
  - 2~3 days revisit time
  - Senses moisture-dependent top surface soil
- **Use cases**

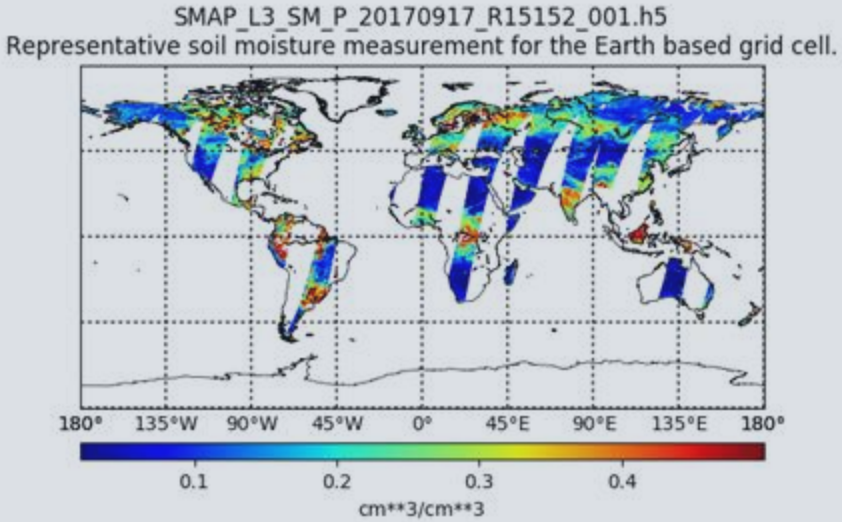
✓ **Weather forecasting (land feedback)**

✓ **Runoff prediction**

✓ **Vegetation water stress**

✓ **Agriculture**

✓ **Inferring precipitation**



Soil moisture prolongation

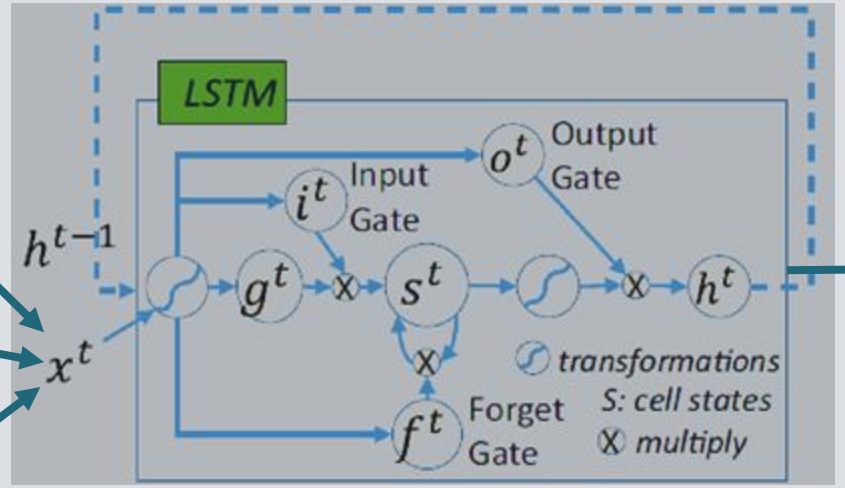
LSTM model

Atm. Forcing (NLDAS)

(optional) Land Surface Model (Noah) solutions

Soil texture, slope, land cover, irrigation, depth to water table, etc

(x)



(y)

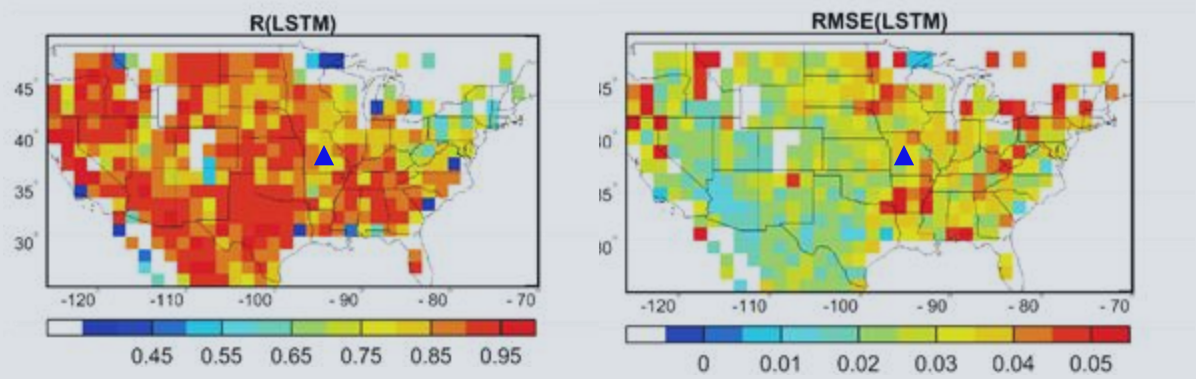
SMAP L3_P

train

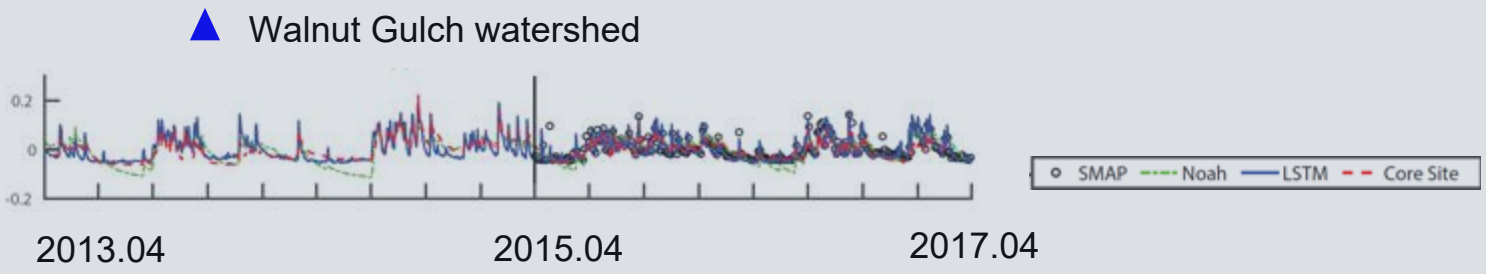
Surface soil moisture

Using LSTM, we prolonged SMAP to …

- **Spatio-temporally seamless coverage**



- **Long-term projections**



Harnessing the scaling effects of big data in geoscientific modeling

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-26107-z OPEN

From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling

Wen-Ping Tsai[1], Dapeng Feng[1], Ming Pan[2,3], Hylke Beck[4], Kathryn Lawson[1,5], Yuan Yang[6,7], Jiangtao Liu[1] & Chaopeng Shen[1,5]

nature reviews earth & environment

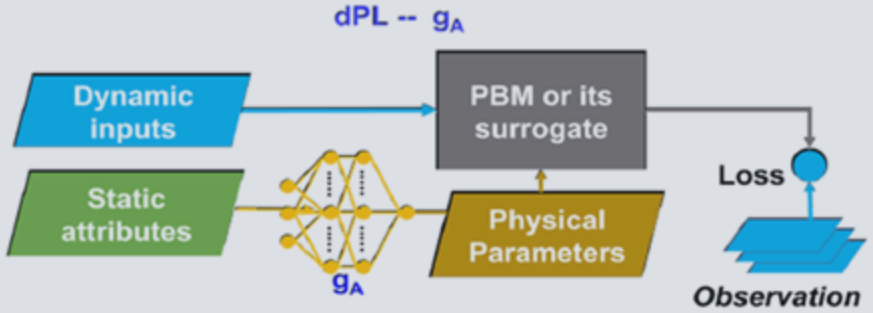
Explore content About the journal Publish with us Subscribe

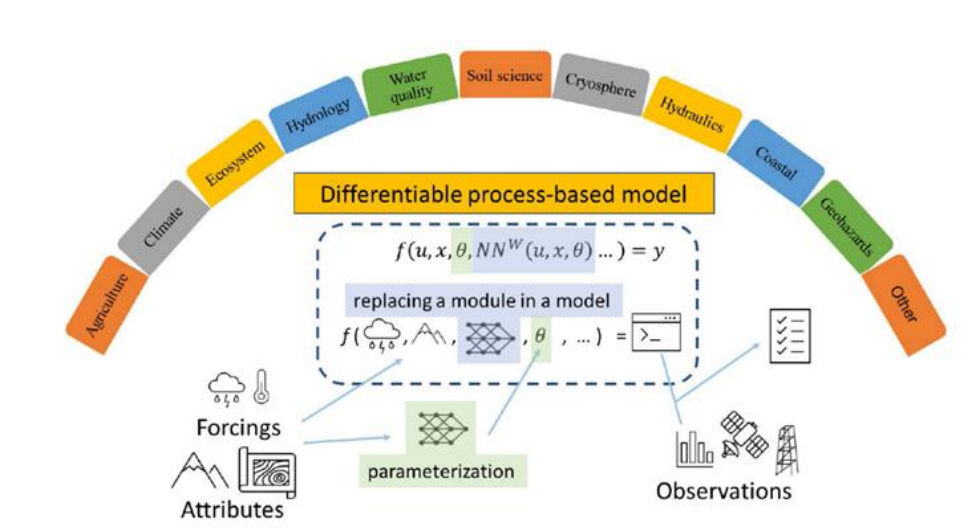
nature › nature reviews earth & environment › perspectives › article

Perspective | Published: 11 July 2023

Differentiable modelling to unify machine learning and physical models for geosciences

differentiable parameter learning (dPL)



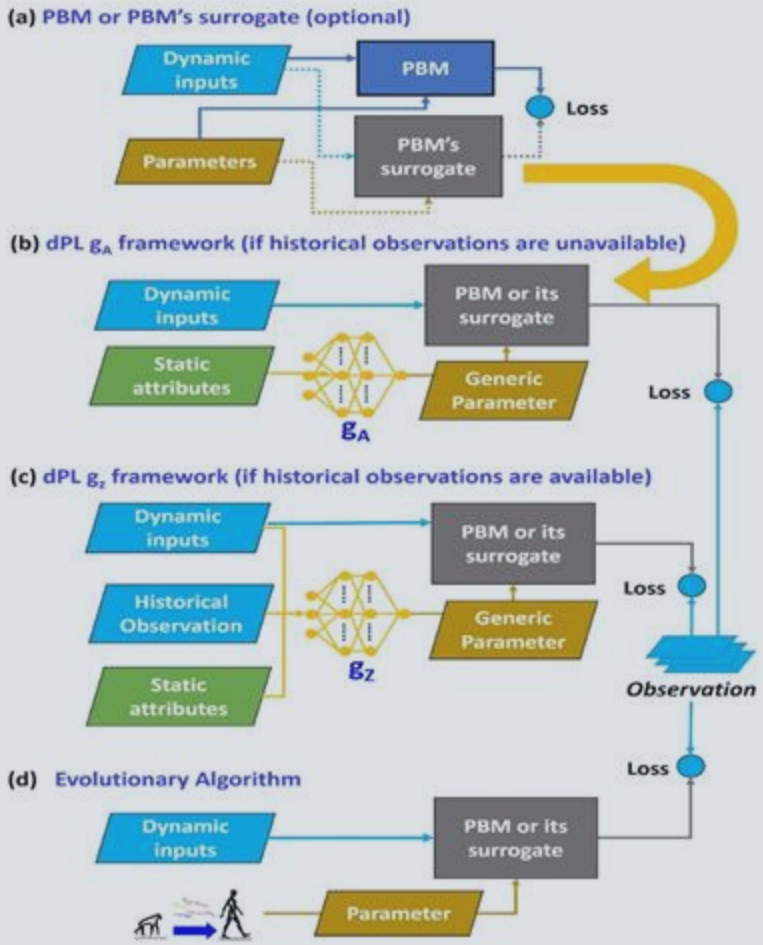


connected the process-based hydrologic model to a neural network that estimates the physical parameters of VIC model

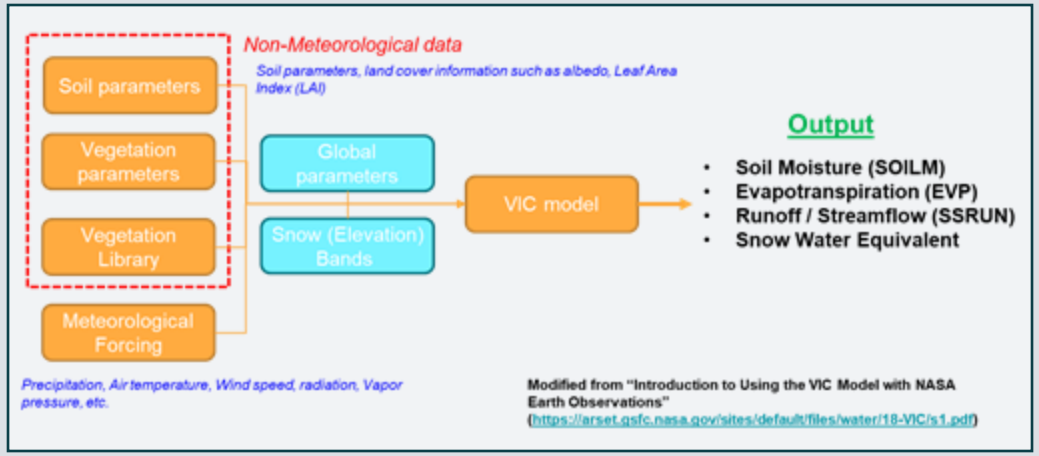
Variable Infiltration Capacity (VIC) Macroscale Hydrologic Model

Tsai et al., 2021, Nature Communications

Differentiable parameter learning (dPL)



Variable Infiltration Capacity (VIC) Macroscale Hydrologic Model

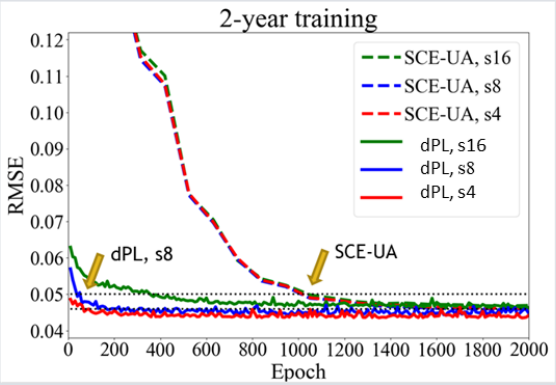


Surrogate model :
Long Short-Term Memory networks (LSTM)

Performance of SMAP soil moisture simulations

(A)

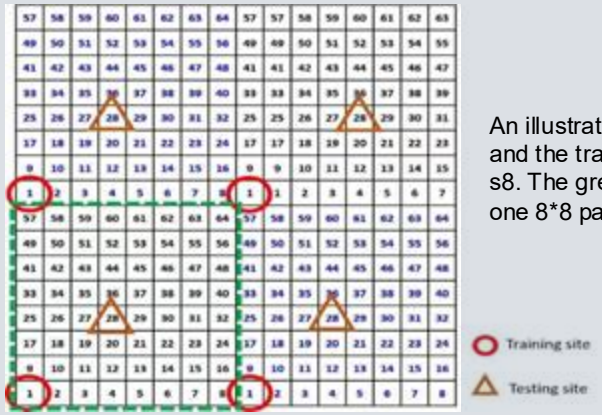
Objective function performance



SCE-UA:
Shuffled Complex Evolution

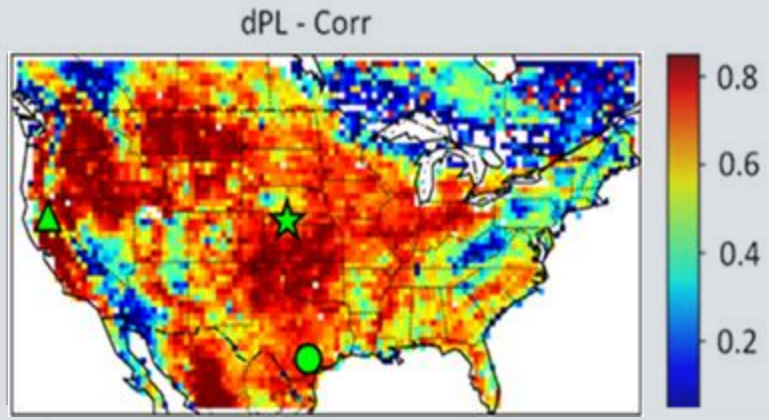
More efficient and scales better with more data!

(B)

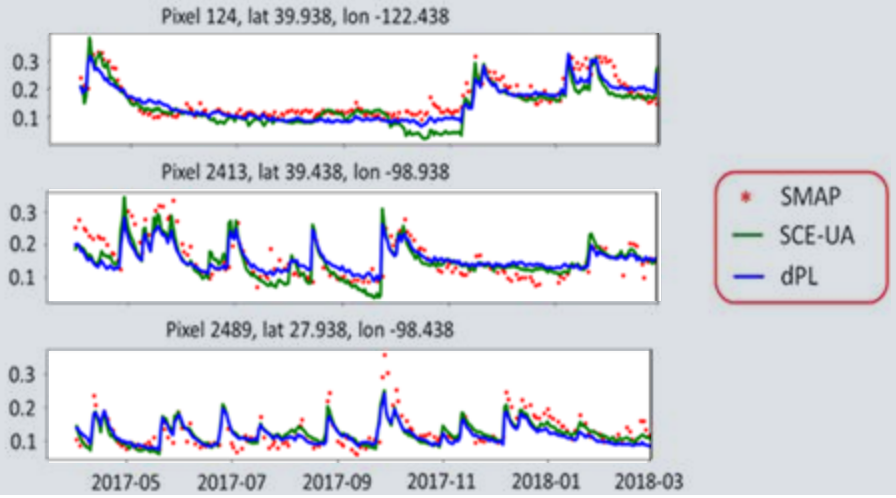


An illustration of the sampling strategy and the training and test sites at density s8. The green dashed line indicates one 8*8 patch (64 gridcells)

(C)

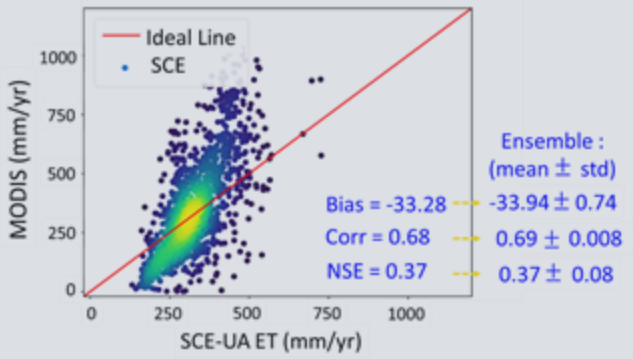


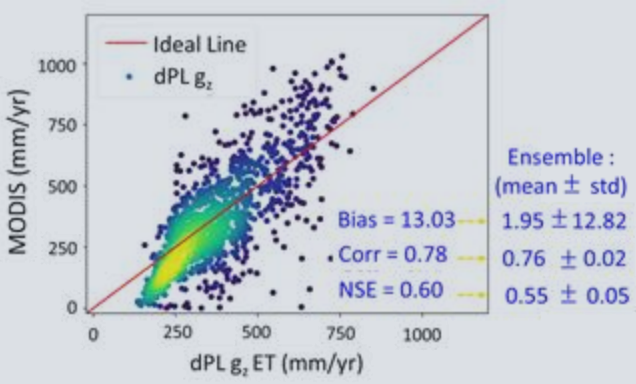
(D)



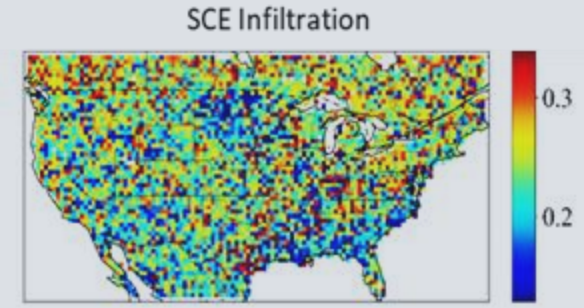
Now capable of modeling other variables such as ET and/or streamflow

(A) Uncalibrated variable: ET prediction

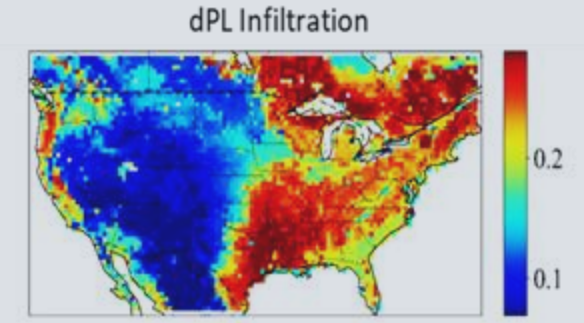




(B) Estimated parameter



jumbled mess

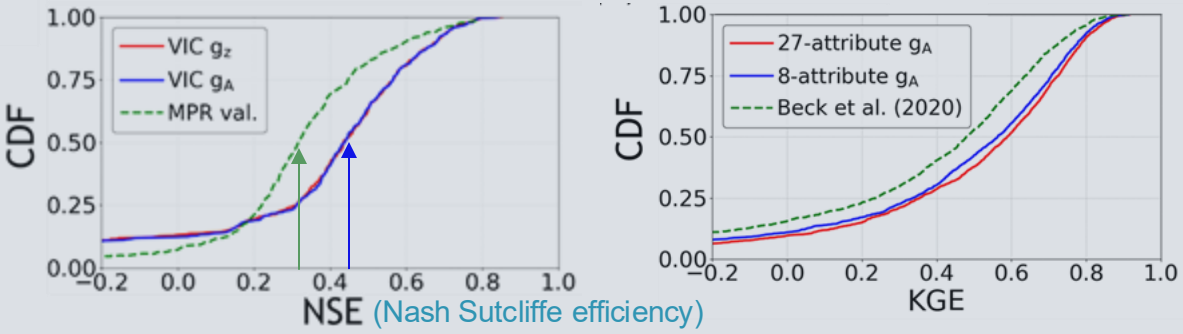


spatial pattern which agrees with our general understanding of physical hydrology and the VIC model's behavior.

Now capable of modeling other variables such as ET and/or streamflow

(C) Streamflow simulation

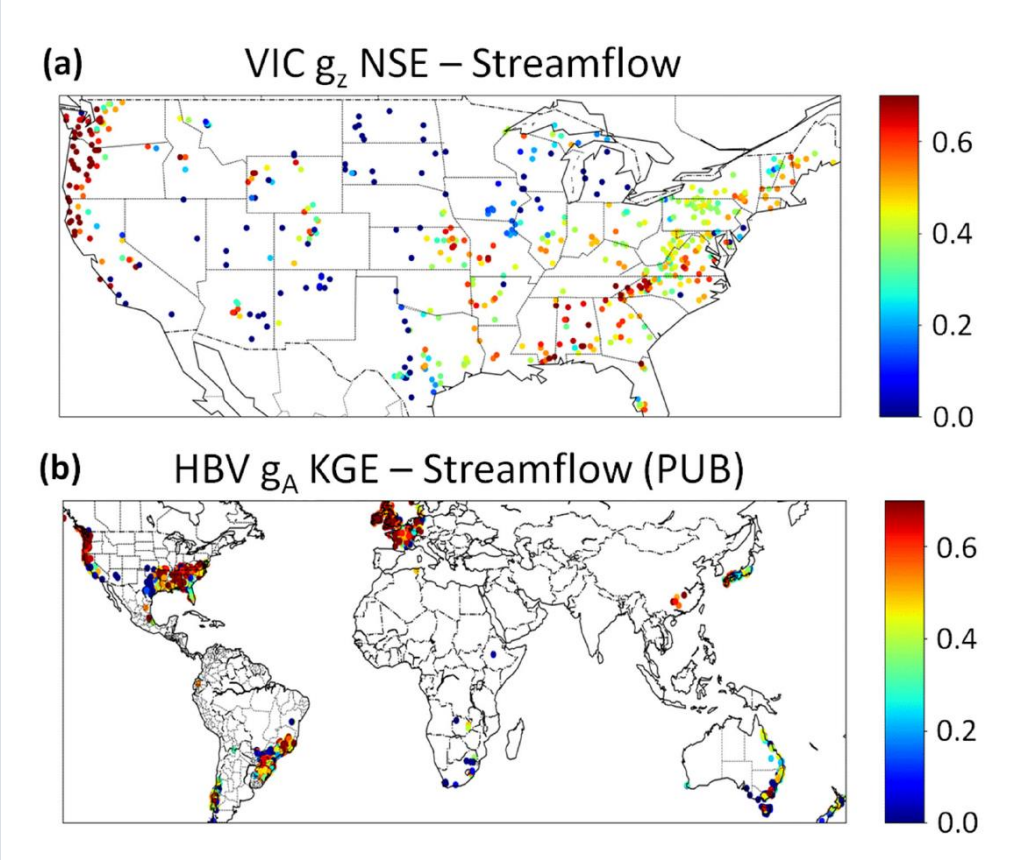
CAMELS (catchment attributes and meteorology for large-sample studies)



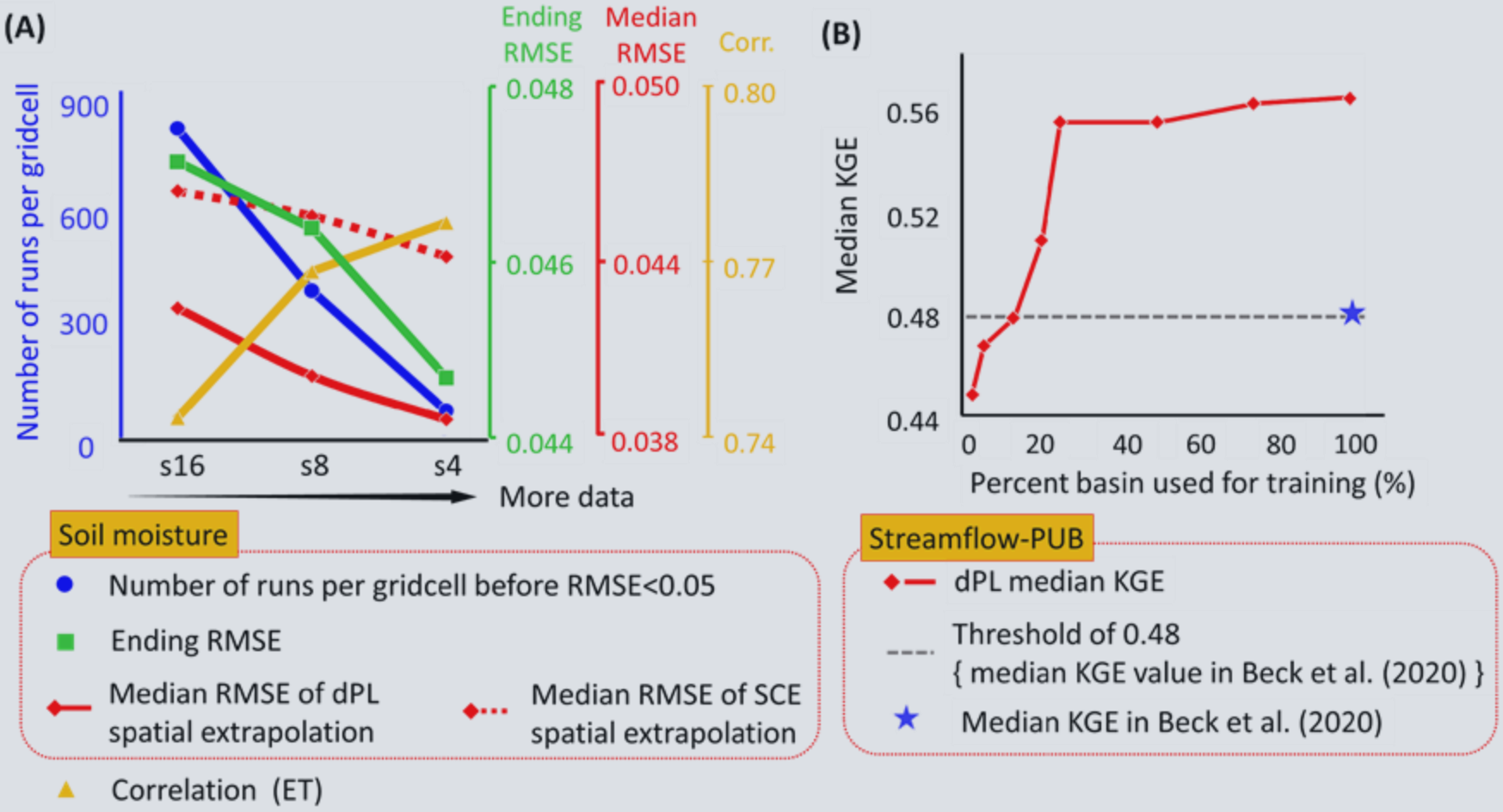
Global case

MPR: Multiscale Parameter Regionalization

MPR過去十多年被廣泛應用於由英國及法國共同開發維護的全球數值天氣預報系統(the European Centre for Medium-Range Weather Forecasts, ECMWF)的綜合預報系統



Scaling curves of dPL



Summary

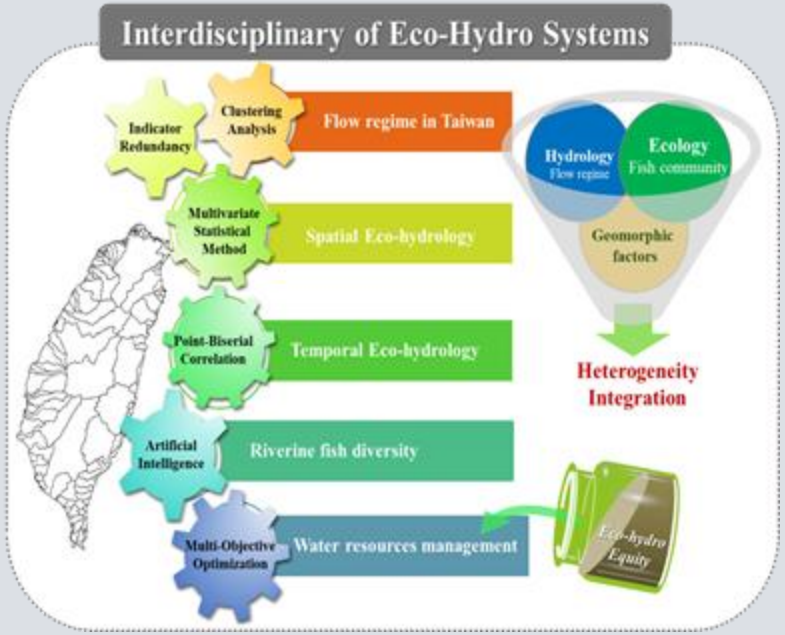
- dPL is stronger than SCE-UA! And More efficient!
- Data from different variables can help improve each other's simulations. Now capable of modeling other variables such as soil moisture, ET, and streamflow.
- As the amount of training data increases, we clearly witness beneficial scaling behaviors, which to our knowledge have never been discussed in the context of geoscientific modeling.

PART.05

Projects

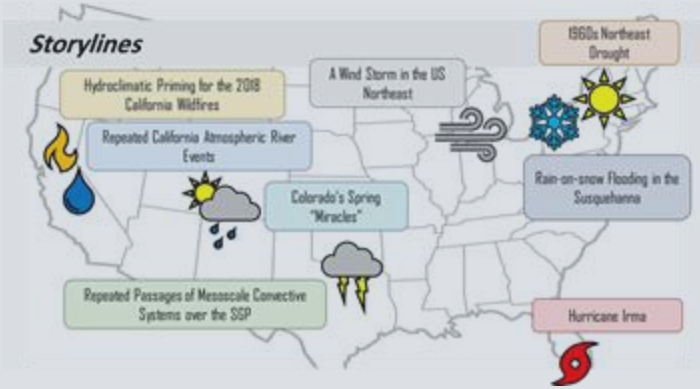
Projects

SUSTAINABLE DEVELOPMENT GOALS
1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL-BEING
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
10 REDUCED INEQUALITIES
11 SUSTAINABLE CITIES AND COMMUNITIES
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
13 CLIMATE ACTION
14 LIFE BELOW WATER
15 LIFE ON LAND
16 PEACE, JUSTICE AND STRONG INSTITUTIONS
17 PARTNERSHIPS FOR THE GOALS
SUSTAINABLE DEVELOPMENT GOALS

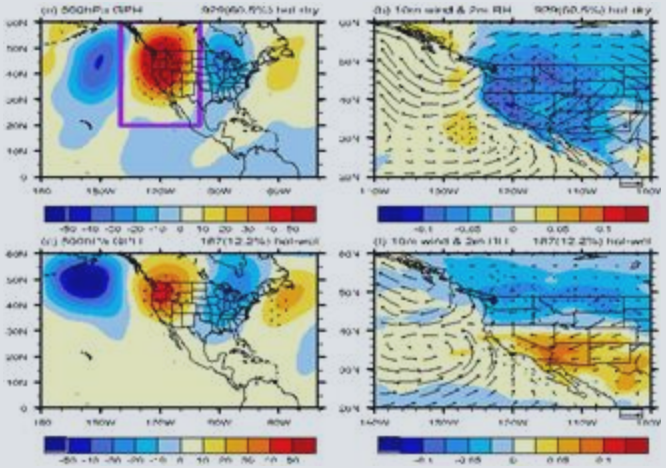

Interdisciplinary of Eco-Hydro Systems
Indicator Redundancy
Clustering Analysis
Flow regime in Taiwan
Multivariate Statistical Method
Spatial Eco-hydrology
Point-Biserial Correlation
Temporal Eco-hydrology
Artificial Intelligence
Riverine fish diversity
Multi-Objective Optimization
Water resources management
Hydrology
Flow regime
Ecology
Fish community
Geomorphic factors
Heterogeneity Integration
Eco-hydro Equity

Increased demand 50% by 2030 (IEA)
Energy
Climate Change
Food
Water
Increased demand 50% by 2030 (FAO)
Increased demand 30% by 2030 (IFPRI)

Projects

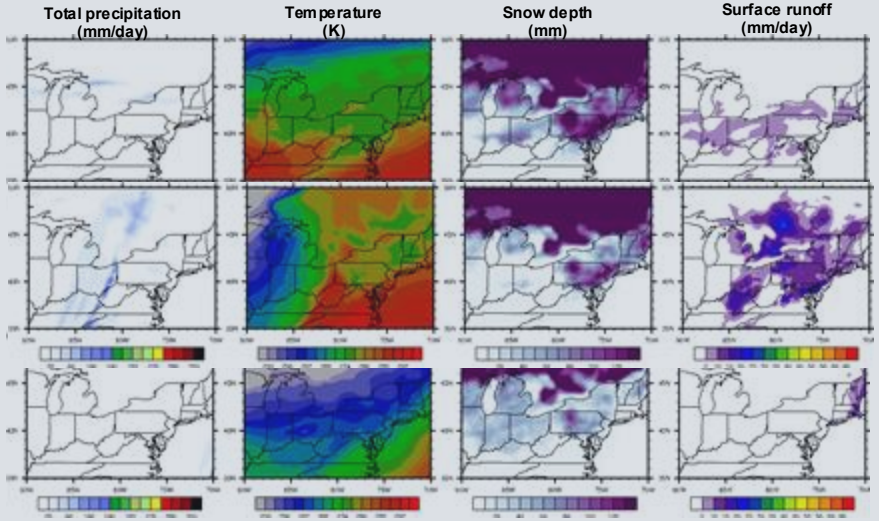


Multi-scale meteorologic model (UCDAVIS)



Basin-scale meteorologic model (PSU)

Study area: Susquehanna river basin



Stop locusts in East Africa now or pay much more to help people later - World Food Programme (WFP)

- With the Food and Agriculture Organization of the U.N. (FAO) support
- Soil moisture prediction model for planning against locust swarms

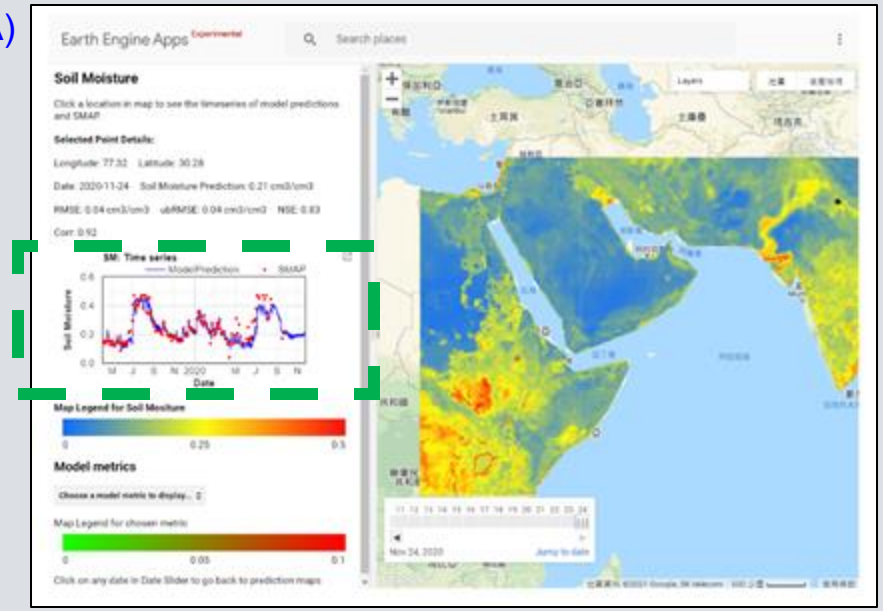
SUNDAY, MARCH 1, 2020

Penn State App Aids UN Efforts To Respond To Crop Destroying Locust Swarms In Africa

By Sara LaJeunesse, Penn State News

Photo: FAO/ Sven Torfinn, Samburu County, Ololokwe, Kenya - A locust swarm takes to the sky in a recent upsurge in northeastern Kenya.

(A)



https://sites.google.com/view/mhpi/locust

deepLDB: a machine-learning-based landslide database and modeling system

This project was one of the 20 grantees of the Google AI Impacts Challenge in 2019, and the mission is to minimize the societal impact of landslide hazards with better predictive capability.

Goals: (1) create a landslide database focusing on events which were not previously reported in the news; and (2) build a model that improves our predictive capability of landslide hazards.

You can two do things on our web portal (i) upload geotiff images or jpeg images and add labels for landslides, or (ii) help add labels of our in-house data.

http://landslide-db-tool.ue.r.appspot.com/

Projects

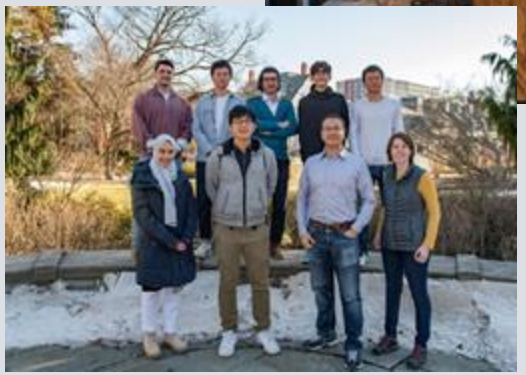
U.S. DEPARTMENT OF ENERGY | Office of Science

HYPERION
UNDERSTANDING HYDROCLIMATE DATA WITH USE-INSPIRED METRICS

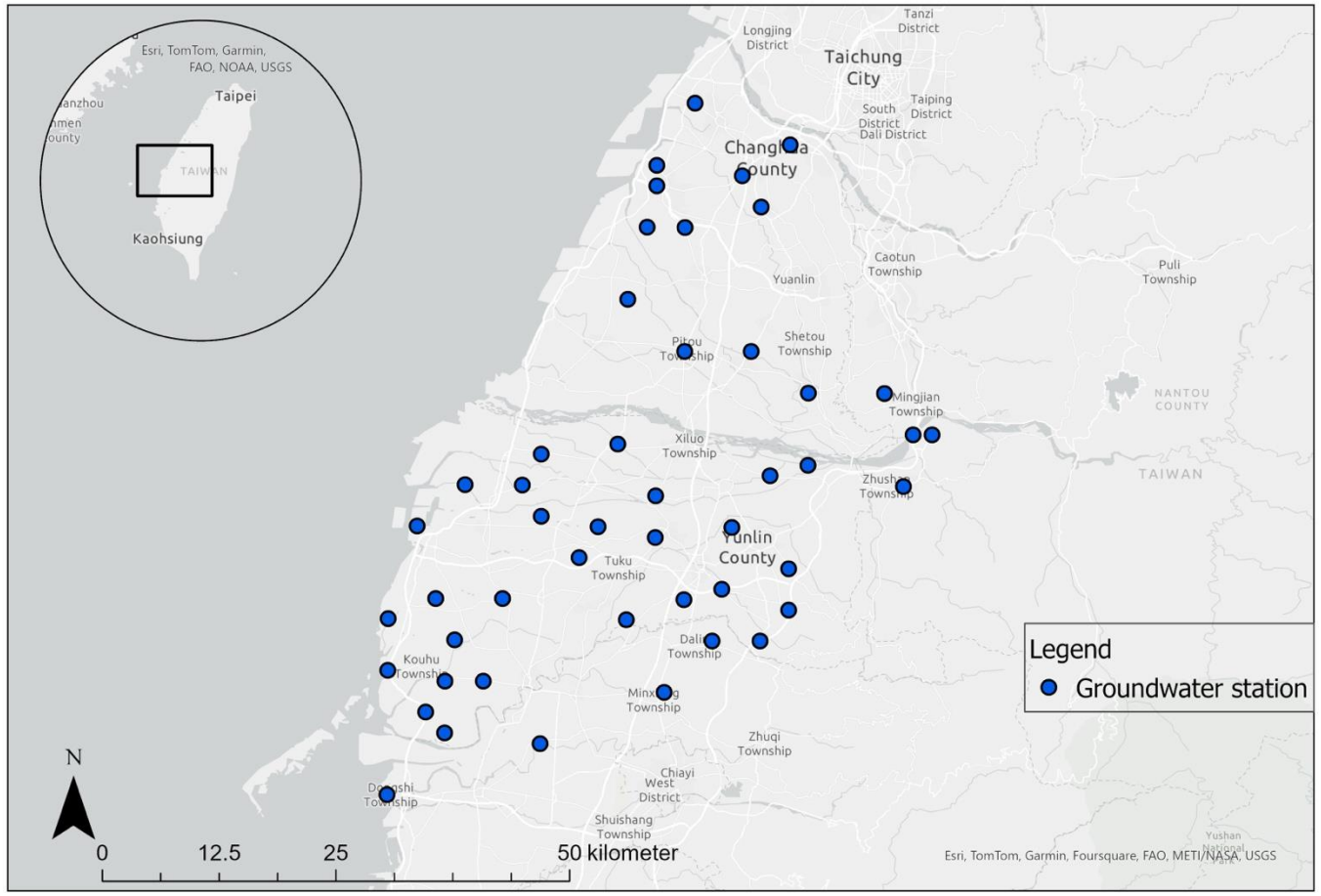
HYPERFACETS
A Framework for Improving Analysis and Modeling of Earth System and Intersectoral Dynamics at Regional Scales

UC DAVIS · Stony Brook University · Cornell University · IOWA STATE UNIVERSITY · Pacific Northwest · UCLA · PennState · cri · UtahState University · NCAR

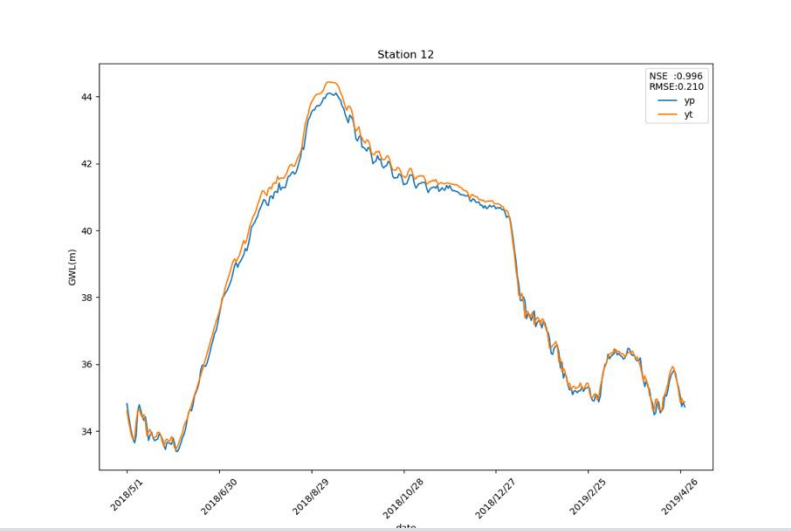
FAO FIAT PANIS



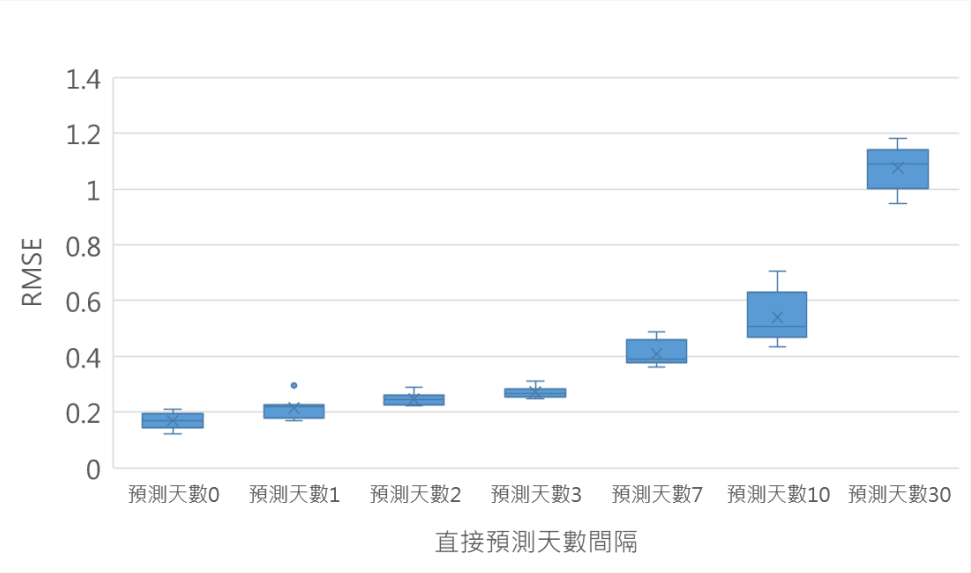
05 中部山區地下水

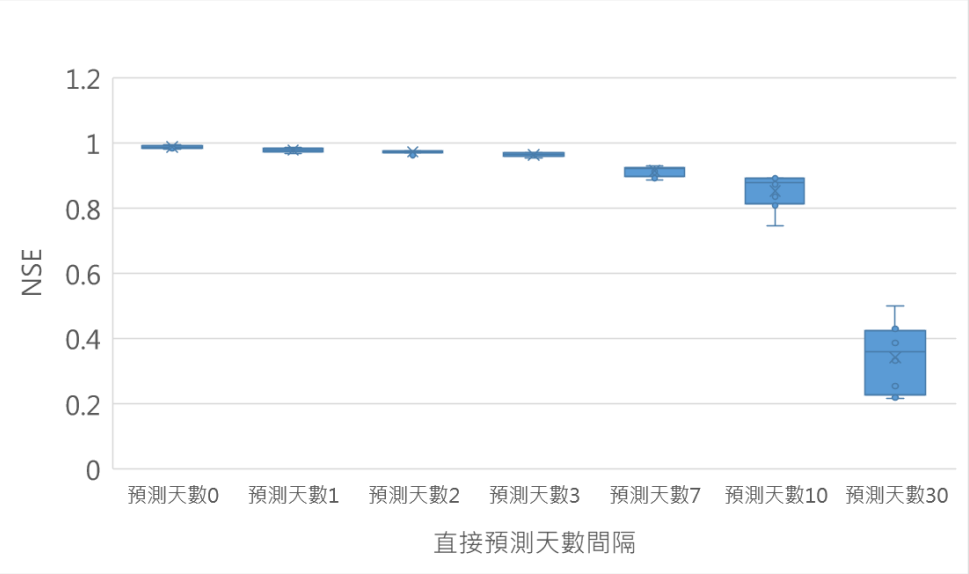


模型各指標中位數	Bias	RMSE	Corr	NSE
基礎模型	1.278	2.152	0.149	-0.626
分區模型	1.423	2.037	0.571	-0.533
資料回歸模型	-0.609	0.749	0.993	0.776
分區&資料回歸模型	-0.002	0.175	0.998	0.989

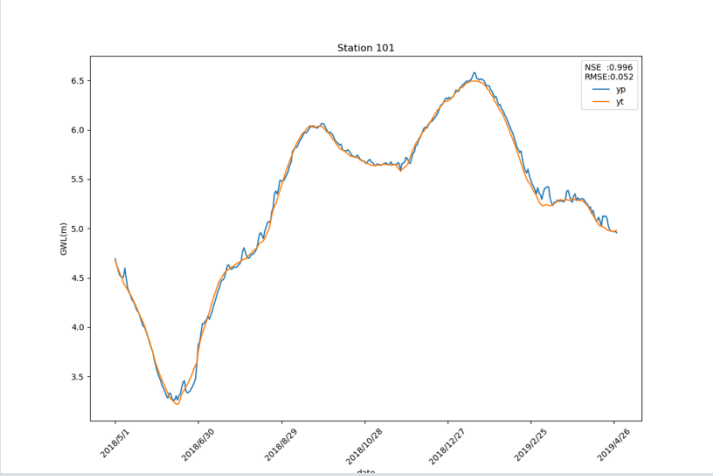


05 未來預測

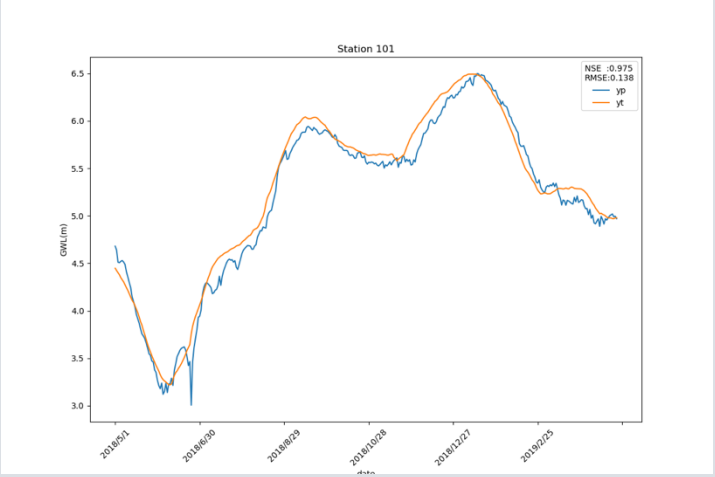




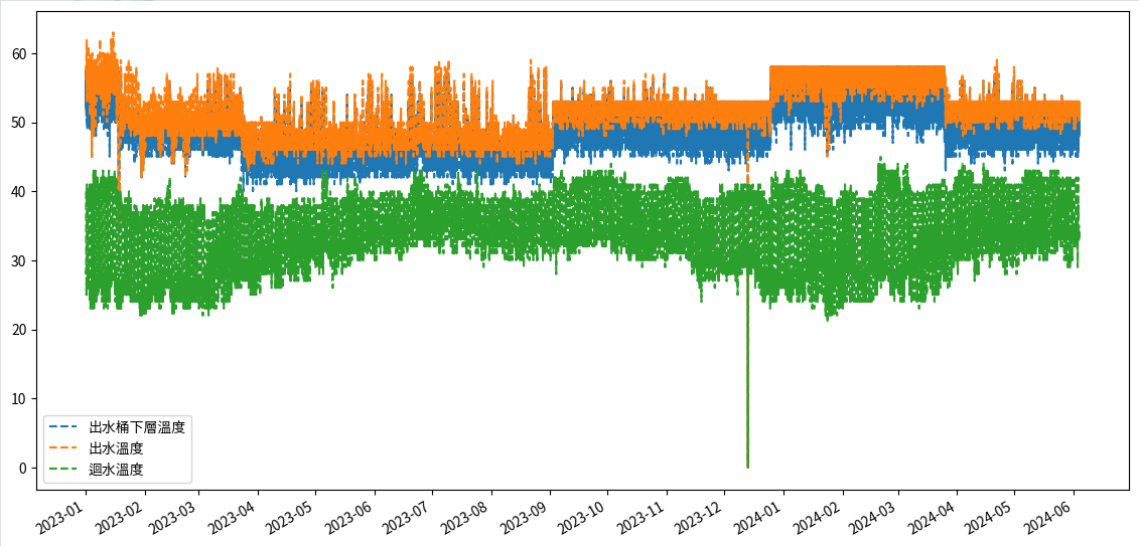
一日後

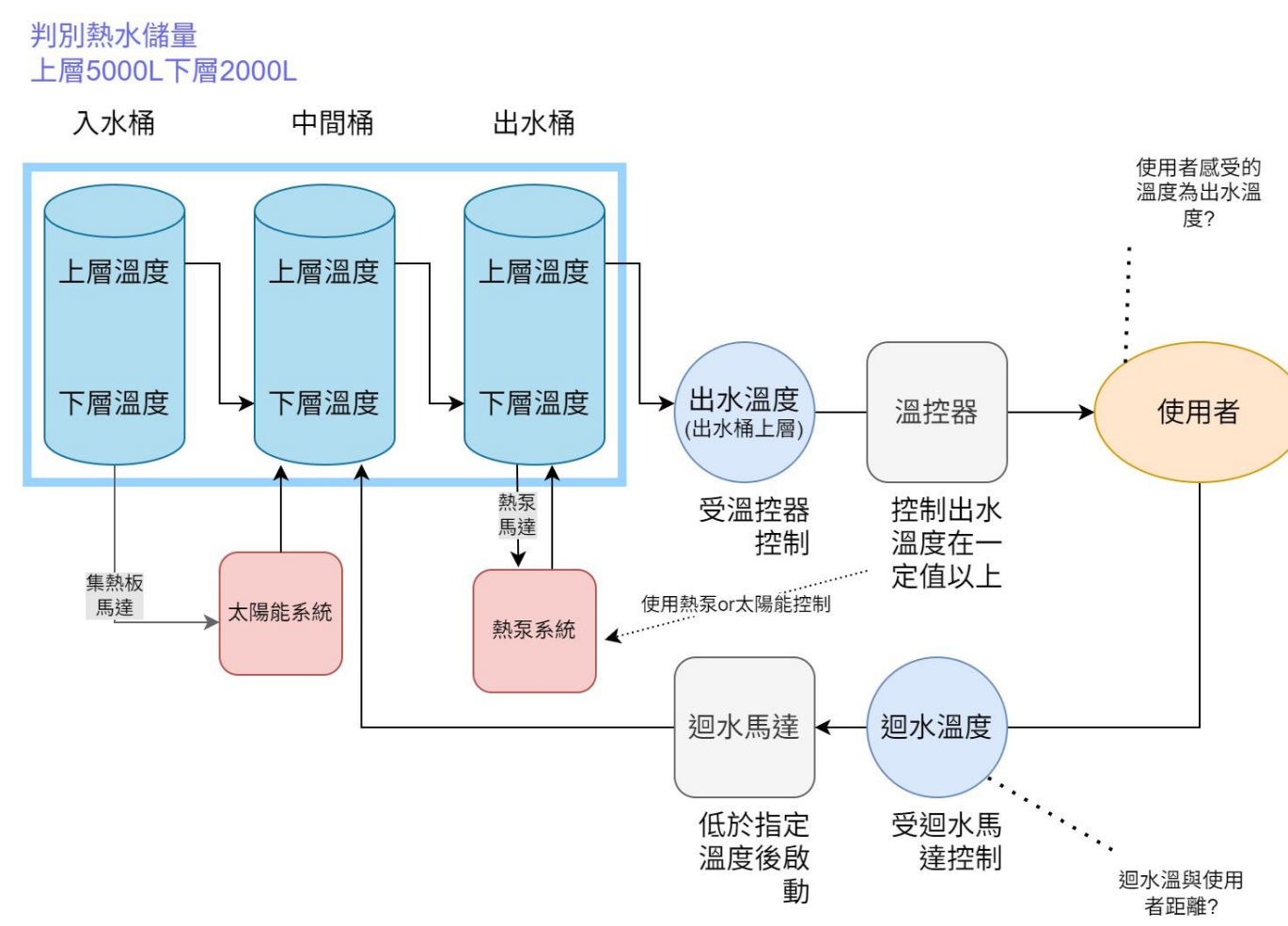


七日後

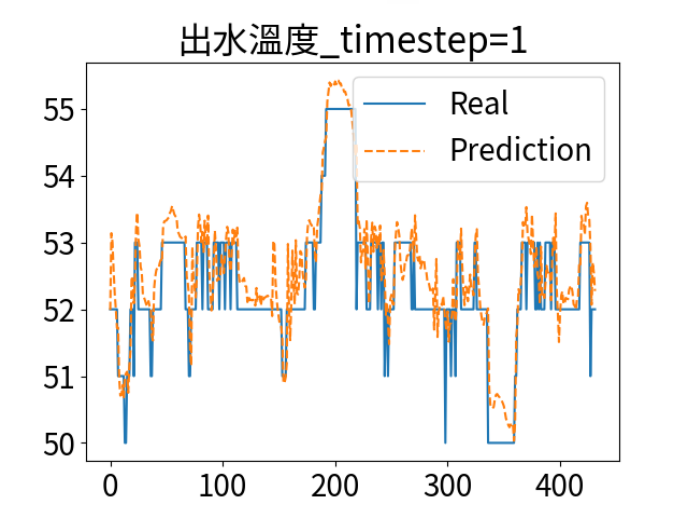


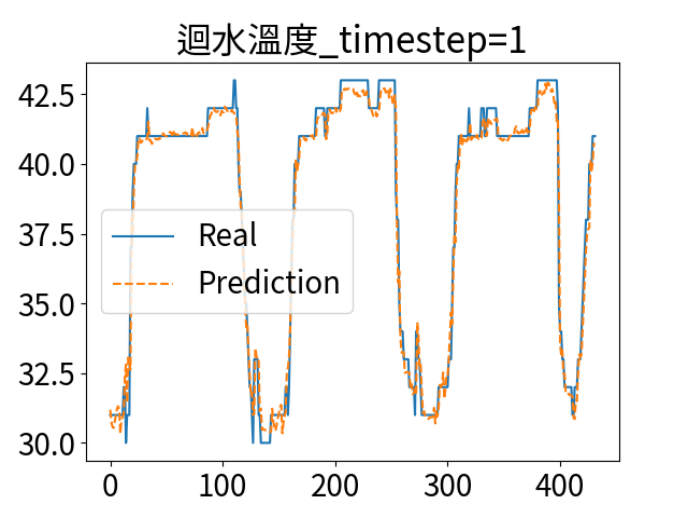
05 水溫預測 -南科宿舍管理局





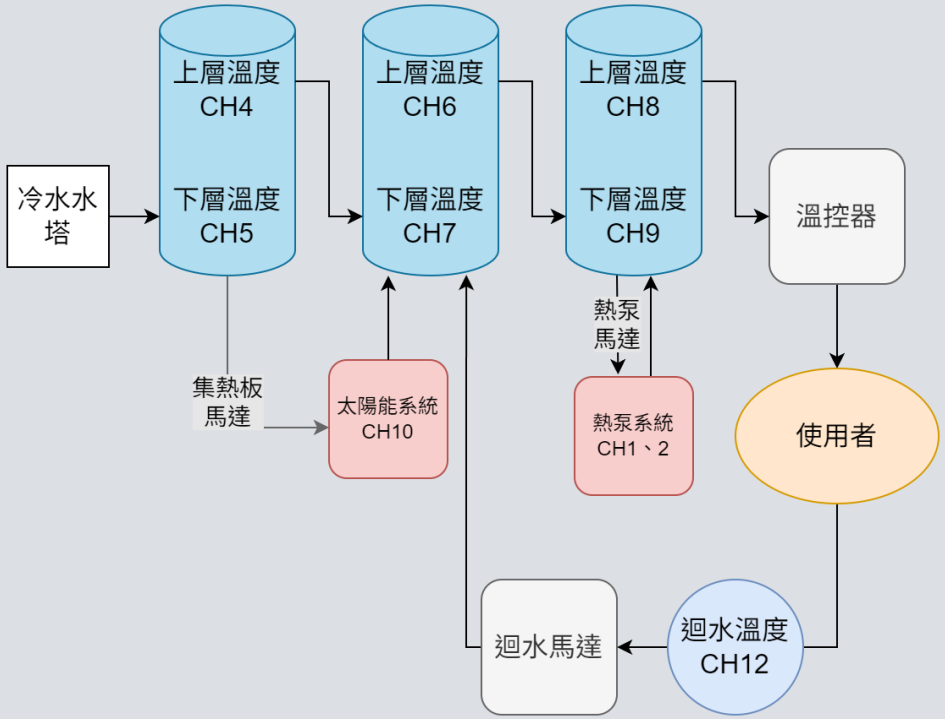
05 水溫預測 -南科宿舍管理局





出水	Corr	RMSE	NSE
10min	0.919	0.583	0.828
20min	0.897	0.641	0.792
30min	0.866	0.725	0.734

迴水	Corr	RMSE	NSE
10min	0.984	0.768	0.967
20min	0.974	0.982	0.946
30min	0.965	1.157	0.925



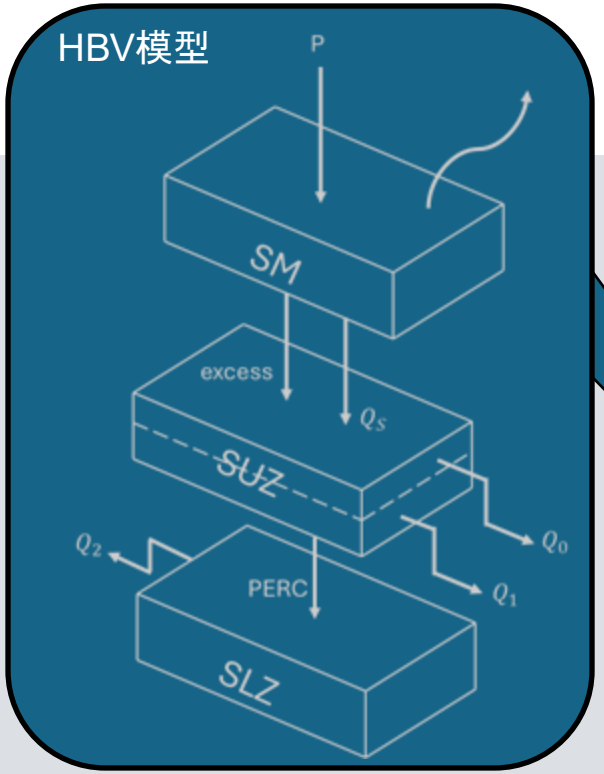
dPL參數學習框架

dPL參數學習框架

藉LSTM參數估計模型生成HBV參數，輸入代理模型生成預測流量。計算之損失函數會回饋給LSTM優化權重，找出最佳的HBV參數。

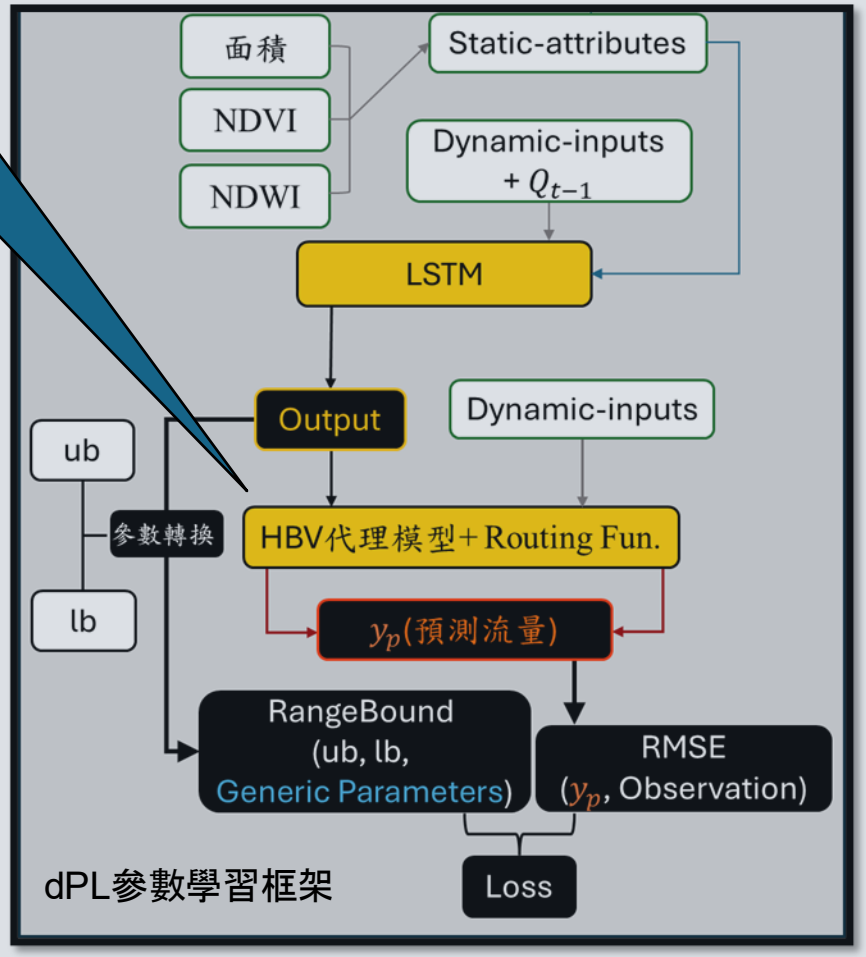
- 輸入資料

動態資料(Dynamic-inputs)
靜態特徵(Static-attributes)



HBV代理模型

HBV模型為降雨-逕流水桶模型，其代理模型使用另一LSTM獨立建置。

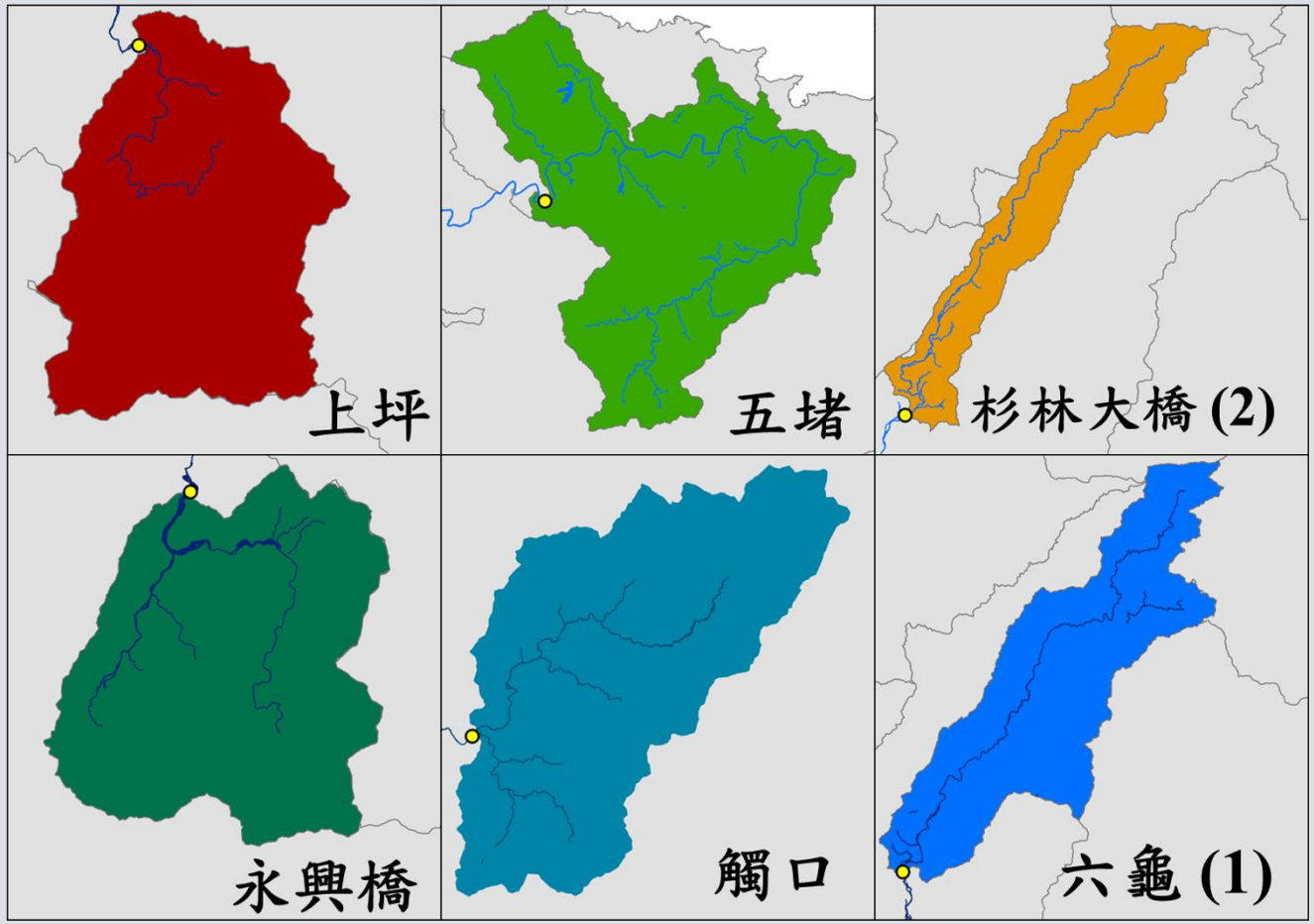


研究區域

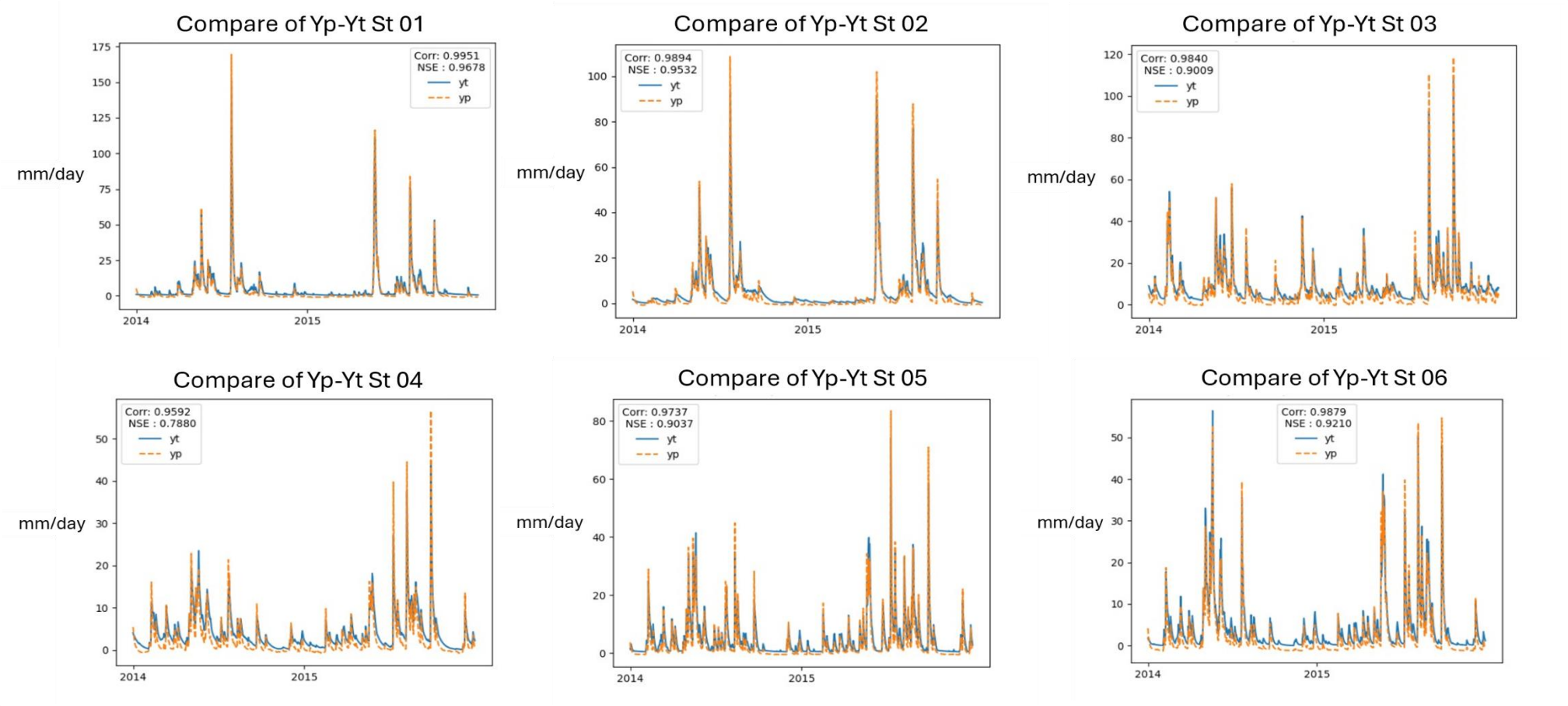
共取6個流量站及其上游集水區作為研究區域:

- 4個流量站位於水庫取水堰上游(五堵、上坪、永興橋、觸口)
- 2個流量站位於高屏溪主要支流(六龜(1)、杉林大橋(2))

流域別	河川名	流量站名	備註
淡水河	基隆河	五堵	新山水庫-八堵抽水站
頭前溪	上坪溪	上坪	寶山第二水庫-上坪攔河堰
中港溪	中港溪	永興橋	永和山水庫-田美攔河堰
八掌溪	八掌溪	觸口	仁義潭水庫-竹山攔河堰
高屏溪	荖濃溪	六龜(1)	
高屏溪	旗山溪	杉林大橋(2)	



代理模型測試


Compare of Yp-Yt St 01
Corr: 0.9951
NSE : 0.9678
yt
yp
mm/day
2014
2015
Compare of Yp-Yt St 02
Corr: 0.9894
NSE : 0.9532
yt
yp
mm/day
2014
2015
Compare of Yp-Yt St 03
Corr: 0.9840
NSE : 0.9009
yt
yp
mm/day
2014
2015
Compare of Yp-Yt St 04
Corr: 0.9592
NSE : 0.7880
yt
yp
mm/day
2014
2015
Compare of Yp-Yt St 05
Corr: 0.9737
NSE : 0.9037
yt
yp
mm/day
2014
2015
Compare of Yp-Yt St 06
Corr: 0.9879
NSE : 0.9210
yt
yp
mm/day
2014
2015

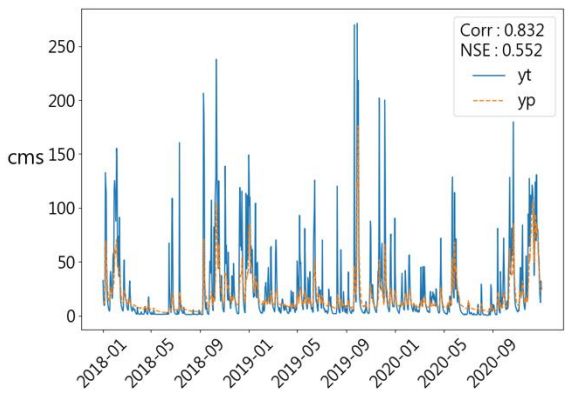
驗 證 成 果

訓練:
2016至2017年，兩年 (日資料)

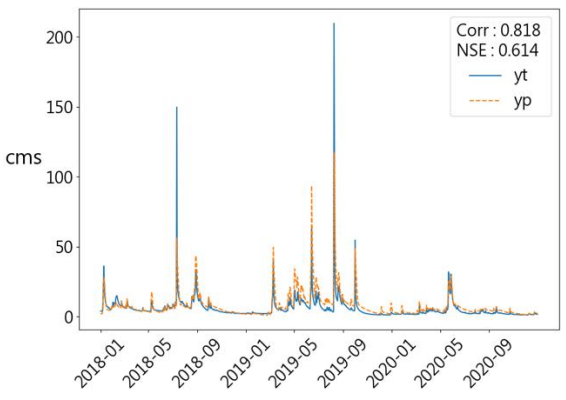
測試:
2018至2020年，三年 (日資料)

藍線 yt => 觀測流量
橘虛線 yp => 模擬流量
Corr => 皮爾遜相關係數
NSE => 納許效率係數

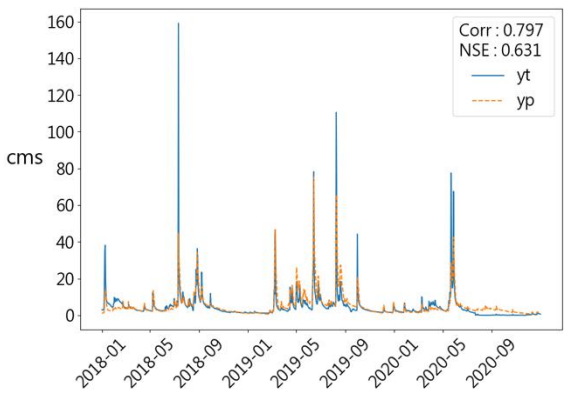
五堵流量站



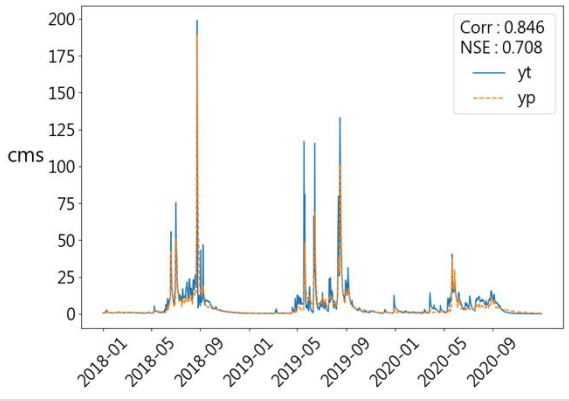
上坪流量站



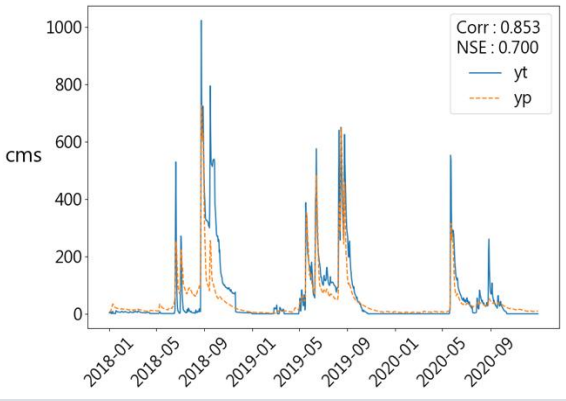
永興橋流量站



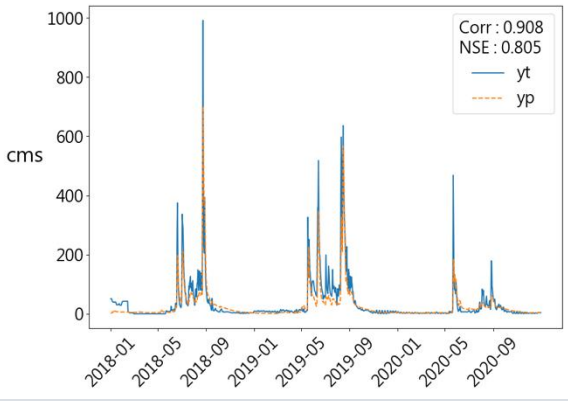
觸口流量站



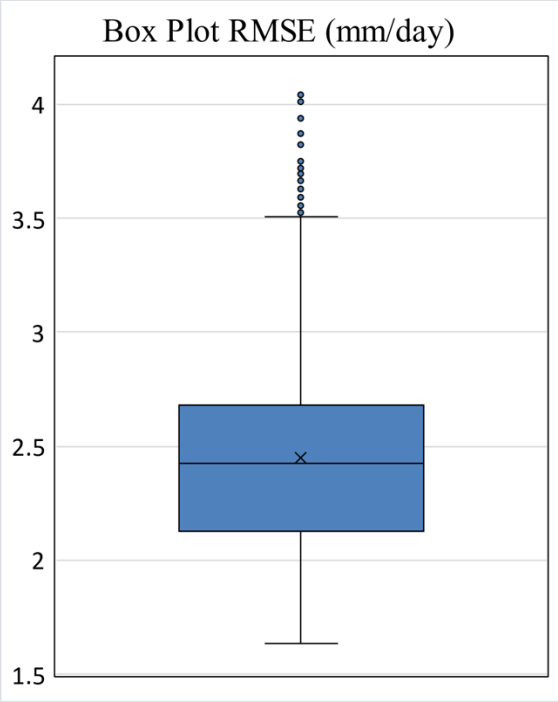
六龜(1)流量站

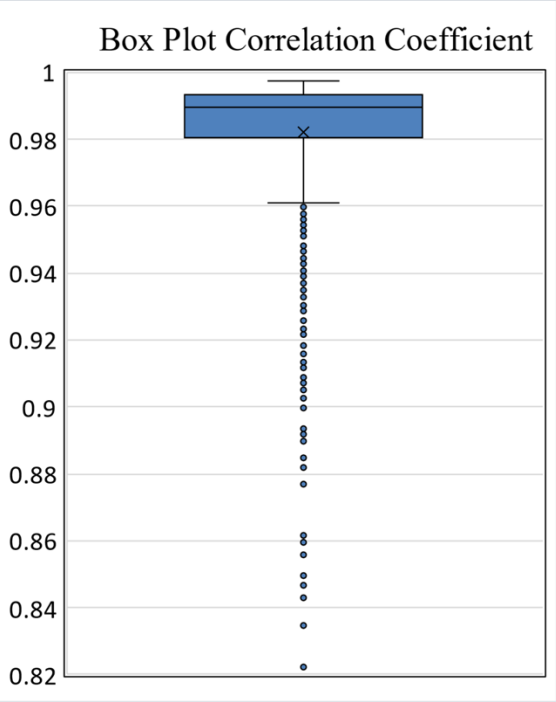


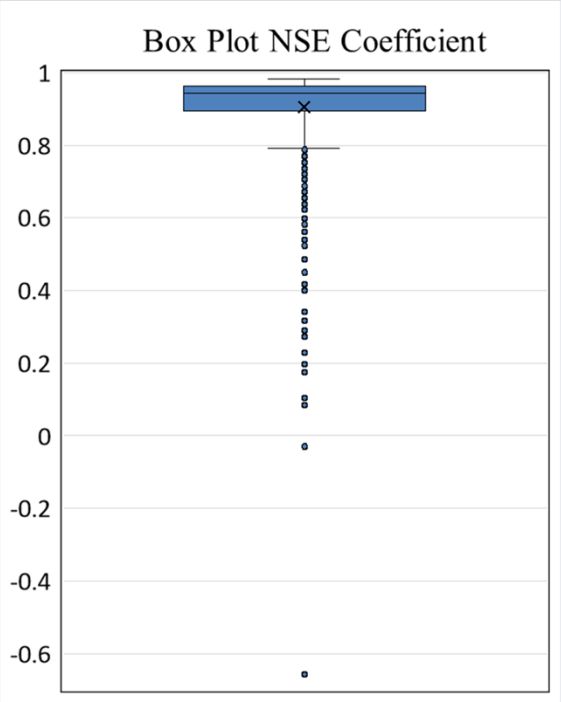
杉林大橋(2)流量站



代理模型測試







	最小值	第一四分位數	中位數	第三四分位數	最大值	平均數
RMSE	1.639	2.130	2.42	2.682	3.510	2.450

	最小值	第一四分位數	中位數	第三四分位數	最大值	平均數
Corr	0.961	0.980	0.990	0.993	0.997	0.982

	最小值	第一四分位數	中位數	第三四分位數	最大值	平均數
NSE	0.791	0.893	0.942	0.962	0.983	0.905

人工智慧與多尺度水文資訊研究室

Thanks for Your Attention