OVERVIEW OF MODERN DATA SCIENCE LANDSCAPE IN HYDROLOGICAL MODELING

GLY606, Aug 28th 2024

University at Buffalo The State University of New York

Learning objectives

- Coding languages?
  - Python? Matlab? Fortran? C++?
  - Where are different languages used in hydrologic models?
  - What is the best practice to learn the coding languages?
- Important concepts
  - Water balance equations
  - What is a watershed?
  - Process-based models
    - Lumped models or distributed models?
  - AI/ML hydrologic models

Coding languages

Python

Matlab

Fortran

C++

More language?

Coding languages

Python

- Python is the most widely used high-productivity language in Scientific Computing.

Matlab

Fortran

C++

More language?

Coding languages

Python

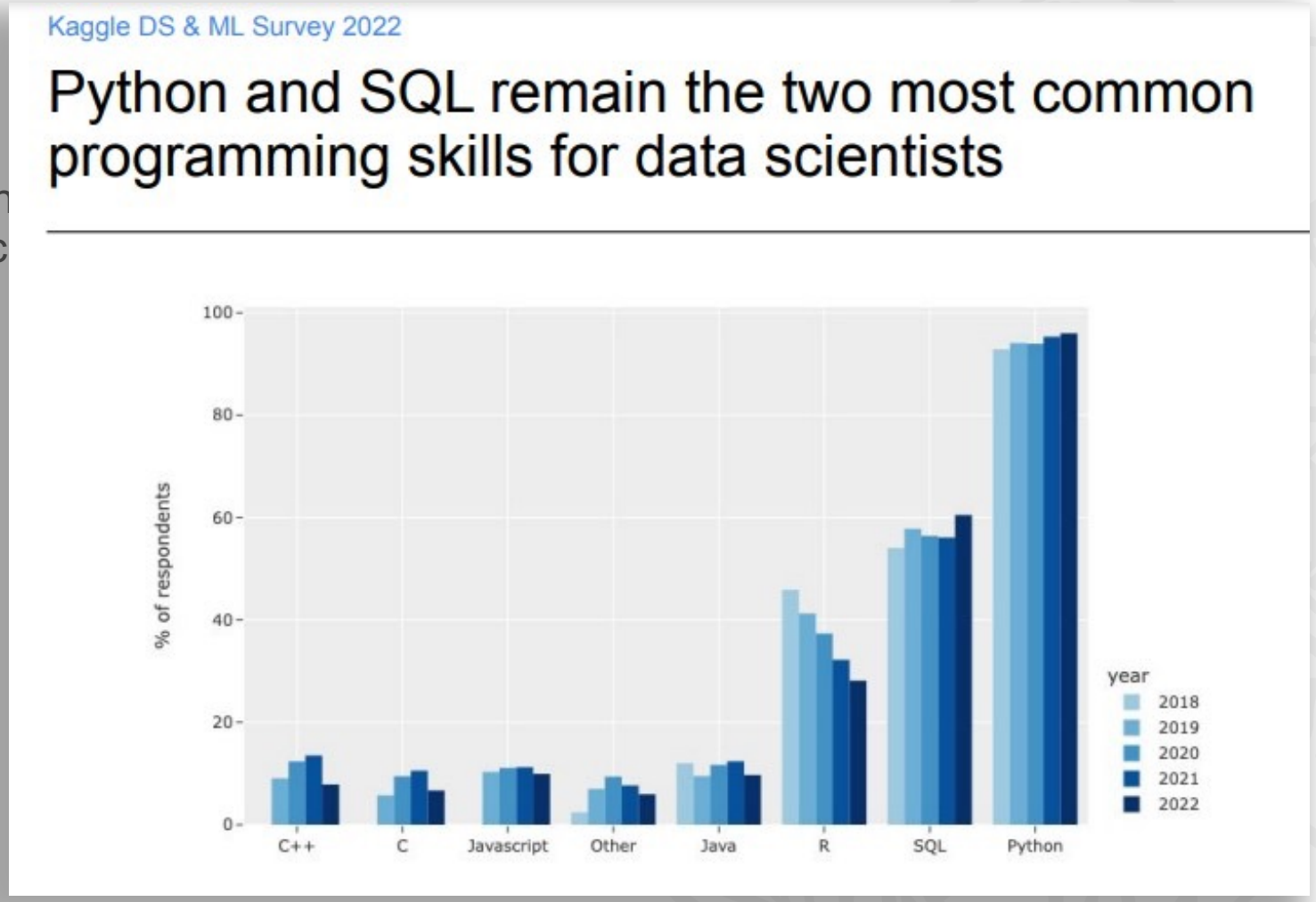
- Python is th
in Scientific

Matlab

Fortran

C++

More language?



Data source: Kaggle DS & ML Survey
Plot credit: https://studyopedia.com/data-science/programming-languages-for-data-science/#google_vignette

Coding languages

Python

Matlab

Fortran

C++

More language?

- Python is the most widely used high-productivity language in Scientific Computing.
- Its very **simple syntax** and broad library support make it ideal for quickly building scalable applications.
- The language does not natively support the type of data structures and other features needed for fast computation, so *few hydrologic models are written in Python.*
- Python is great for data analysis so we will mainly use Python in this class!

Coding languages

Python

Matlab

Fortran

C++

More language?

- Matlab is one of the oldest high-productivity languages and has been the defacto standard for fast numerical prototyping before Python.
- It is still heavily used in many numerical applications, given its excellent toolbox and huge amount of legacy code that exists.

Coding languages

Python

Matlab

Fortran

C++

More language?

- Matlab is one of the oldest high-productivity languages and has been the defacto standard for fast numerical prototyping before Python.
- It is still heavily used in many numerical applications, given its excellent toolbox and huge amount of legacy code that exists.
- **Paywall!!!** While Matlab has quite favorable licenses for academic use, it is expensive for commercial use, and if possible Python as open-source alternative is preferable for new projects.

Coding languages

Python

Matlab

Fortran

C++

More language?

- Fortran is one of the dinosaurs of scientific computing. Fortran originated in the 1950s and its most recent incarnation is Fortran 2018.
- Fortran is still actively used for a lot of HPC code, especially when it comes to legacy applications.
- Many Global Climate Models and hydrologic models were written in Fortran!

Coding languages

Python

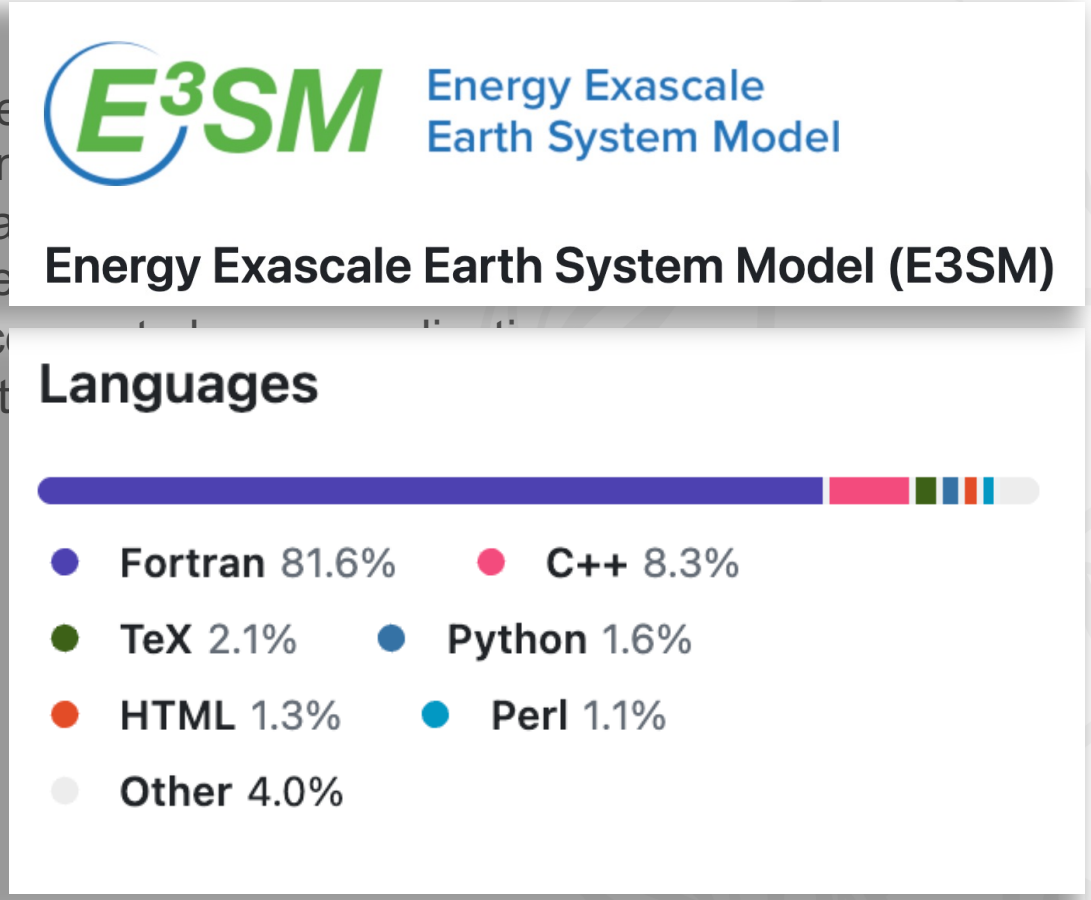
Matlab

Fortran

C++

More language?

- Fortran is one of the
  Fortran originated ir
  incarnation is Fortra
- Fortran is still active
  especially when it c
- Many Global Climat
  written in Fortran!



Source: https://github.com/E3SM-Project/E3SM (accessed on Aug 26th 2024)

Coding languages

Python

Matlab

Fortran

C++

More language?

- C++ is the default language of Scientific Computing.
- It is mature, has a huge ecosystem and most modern heterogeneous compute environments (Cuda/Sycl, etc.) are developed for C++.
- Similar to Fortran, it requires compilation and it is not as user-friendly as Python or Matlab, so it has a higher learning curves.

Examples

- Assign variable ***a*** with the value 1
- Assign variable ***b*** with the value 2
- Calculate the sum of ***a*** and ***b*** and assign the value to variable ***result***
- Print ***result***

https://www.leetcode.com

The best ways to learn coding is through solving problems

Hydrologic models

Hydrology 101

Water balance equation

P =

Precipitation

Hydrologic models

Hydrology 101

Water balance equation

$$P = Q + E + \Delta S$$

Runoff

Storage change

Precipitation

Evapotranspiration

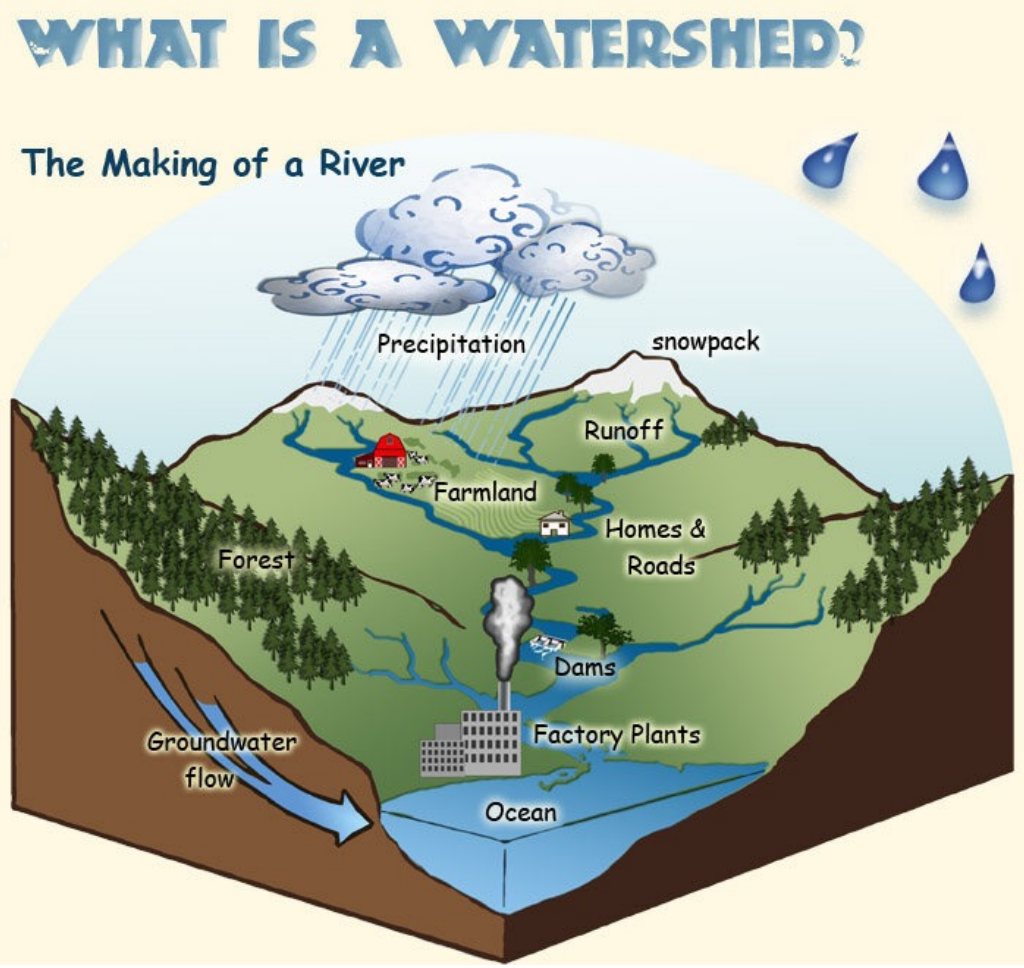
Hydrologic models

Hydrology 101

Water balance equation

Runoff is the main variable of interest to hydrologist!

Watershed

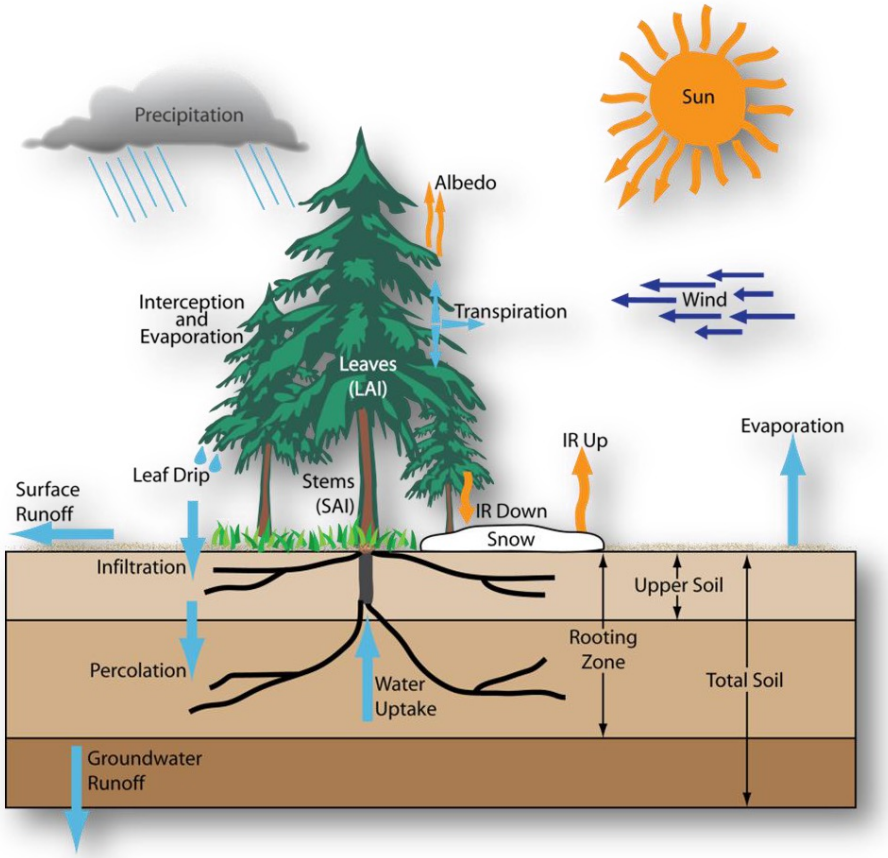


- **Watershed** describes an area of land that drains downslope to the lowest point.
  - Imagine a water drop falls on a mountain: where will it flow?
- Watershed boundaries follow major ridgelines around channels and meet at the bottom.
- Watersheds can be large or small.

What watershed are we located at?

- United States Geological Survey (USGS) National Watershed Boundary Dataset
  - Buffalo (https://hub.arcgis.com/maps/esri::usgs-watershed-boundaries/explore?location=42.752919%2C-78.410536%2C8.05)

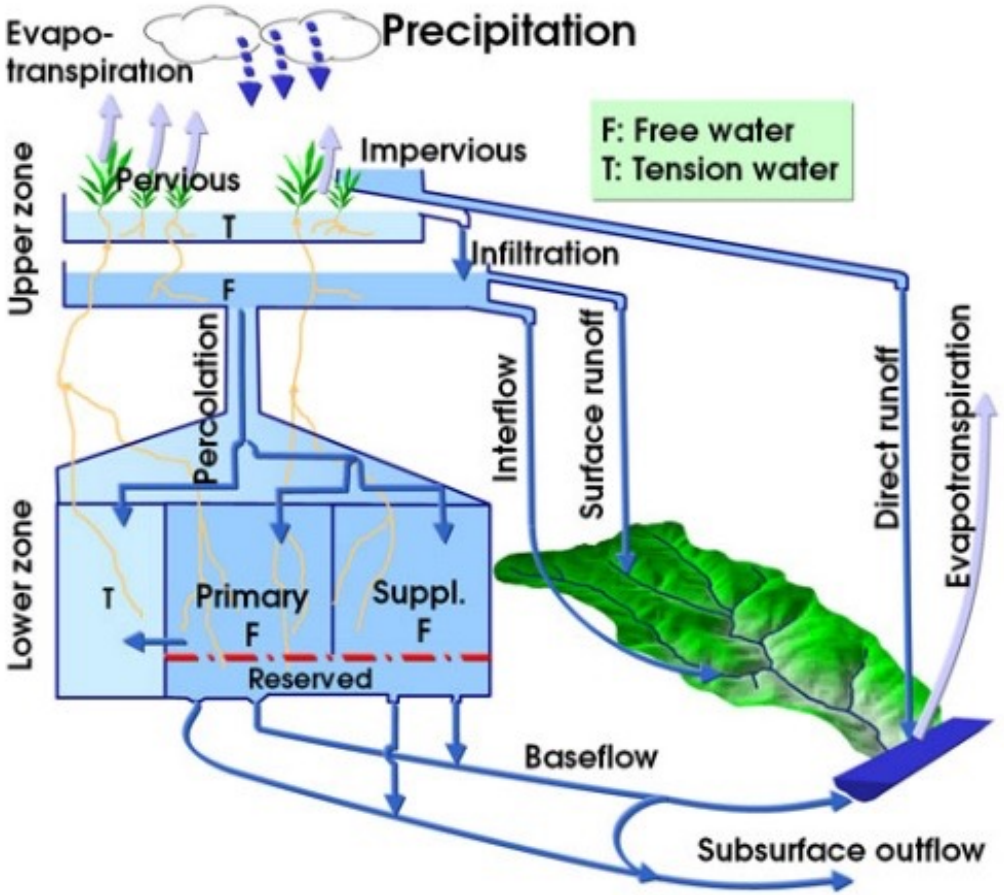
Process-based hydrologic models



https://regclim.coas.oregonstate.edu/teaching-examples/climatology-201/energy-and-water-balance-background/

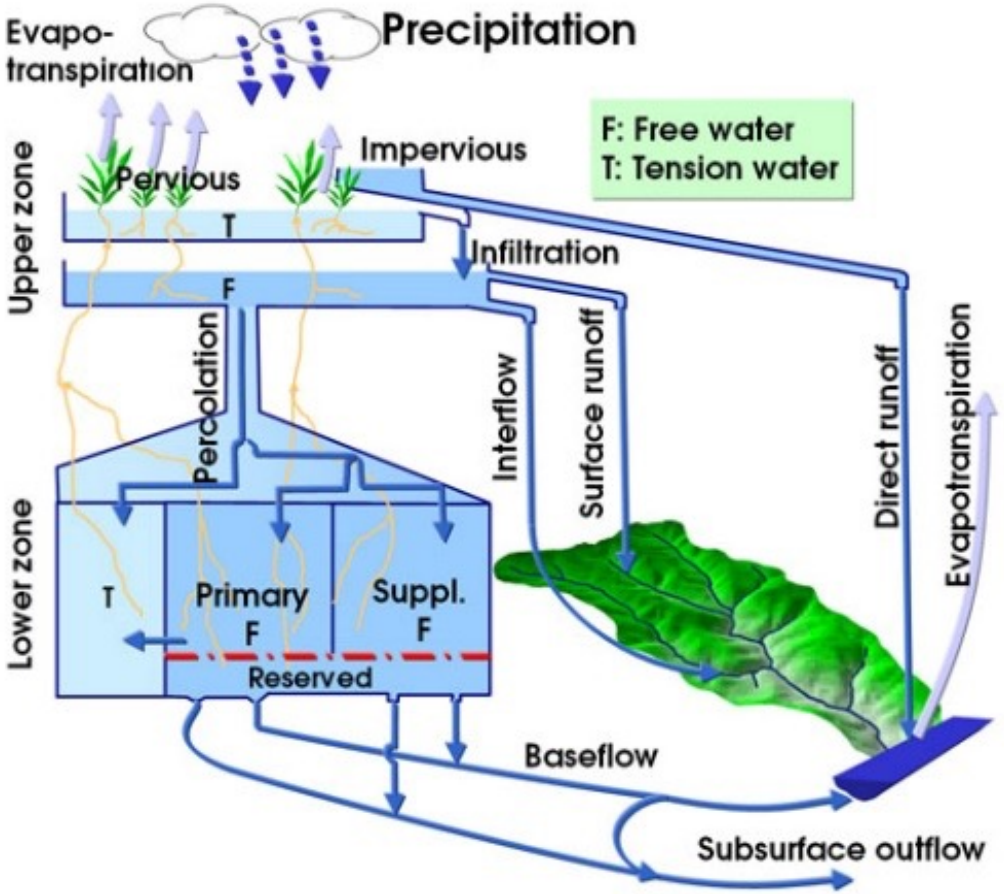
- Process-based hydrologic models represents a collection of connected processes, such as soil infiltration, soil evaporation, transpiration from vegetation, etc.
- Closure of water balance and energy balance
- Complexities of models (different perspectives)
  - Lump model or spatially distributed models
  - Physical process representation

Sacramento Soil Moisture Accounting Model (SAC-SMA)



- SAC-SMA is a lumped hydrologic model
  - Newer development might enable it to be semi-distributed.
- The history of model development goes back to 1970s.
- Highly abstraction of physical processes related to real-world water cycles
- It is probably one of the most famous and widely used hydrological models

Sacramento Soil Moisture Accounting Model (SAC-SMA)

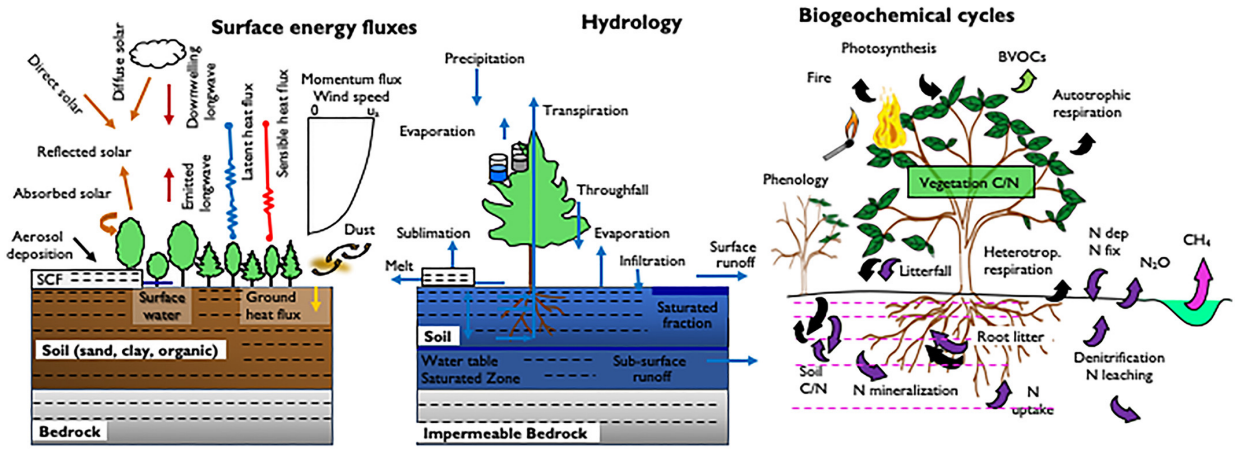


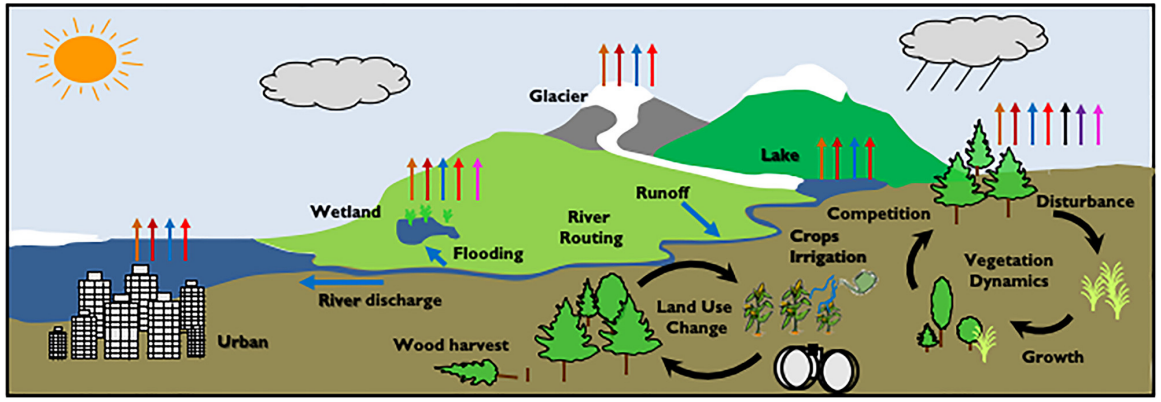
Languages https://github.com/NOAA-OWP/sac-sma

- Fortran 95.5%
- CMake 2.3%
- Makefile 2.2%

- SAC-SMA is a lumped hydrologic model
  - Newer development might enable it to be semi-distributed.
- The history of model development goes back to 1970s.
- Highly abstraction of physical processes related to real-world water cycles
- It is probably one of the most famous and widely used hydrological models

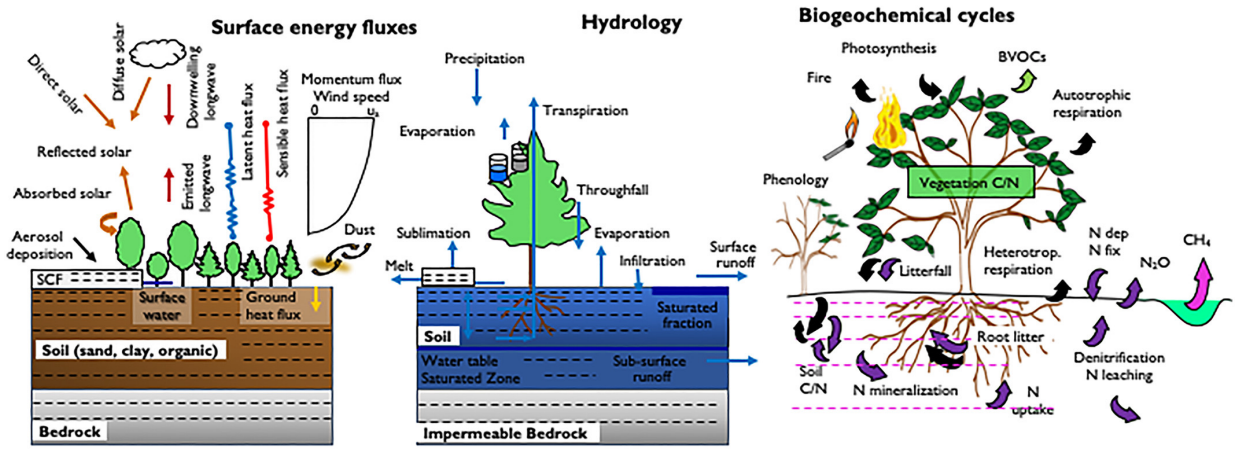
Community Terrestrial Systems Model (CTSM)

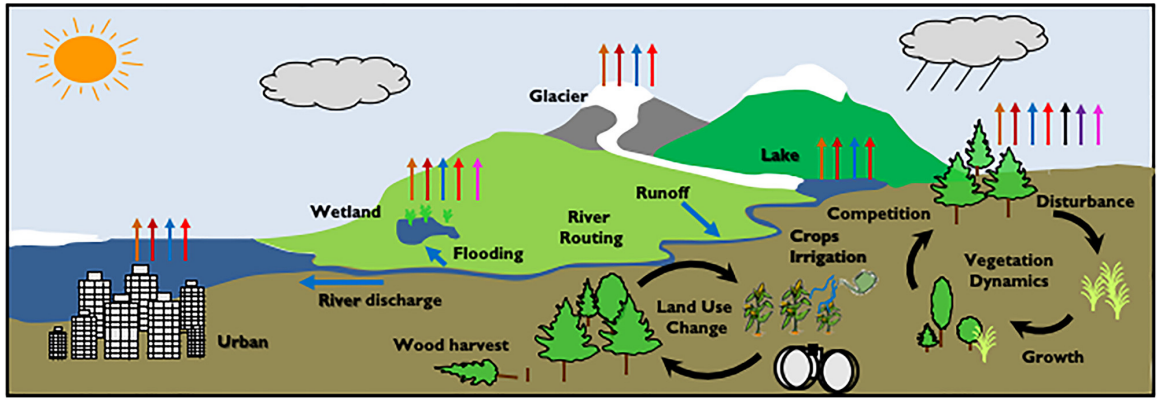




- CTSM is a distributed hydrologic model
- The history of model development goes back to 1996.
- State-of-the-science land models that more closely mimic the real-world physical processes, not only for water but energy and biogeochemical cycles.
- It is widely used in earth system modeling community.

Community Terrestrial Systems Model (CTSM)





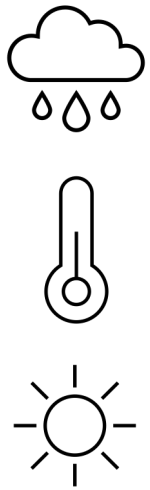
Languages

- Fortran 86.0%
- Python 9.0%
- Perl 3.1%
- Shell 0.7%
- NCL 0.6%
- CMake 0.4%
- Other 0.2%

- CTSM is a distributed hydrologic model
- The history of model development goes back to 1996.
- State-of-the-science land models that more closely mimic the real-world physical processes, not only for water but energy and biogeochemical cycles.
- It is widely used in earth system modeling community.

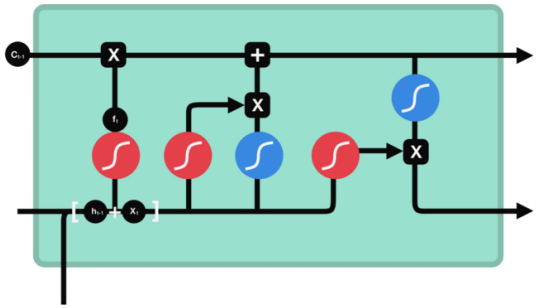
ML hydrologic models

Input



Meteorological forcing data

ML-model

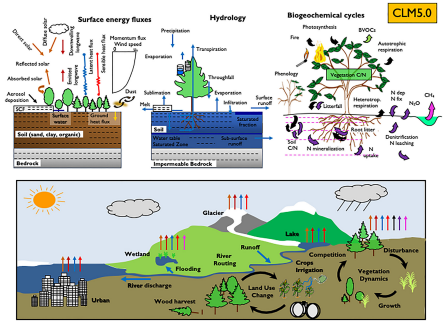


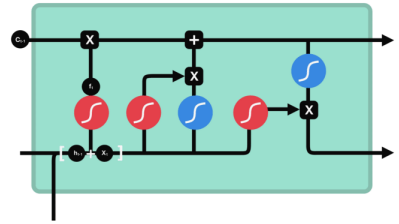
Output

River flow

- Data-driven model
- Directly used meteorological forcing data to predict runoff
  - Detailed physical processes are usually not explicitly represented
- Black box nature
- Water balance and energy balance are not explicitly represented.

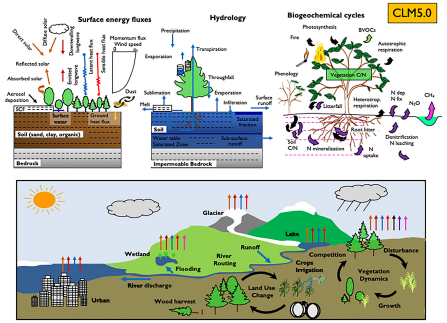
It is not just black or white!





Pure process-based models	Pure ML-AI models

It is not just black or white!

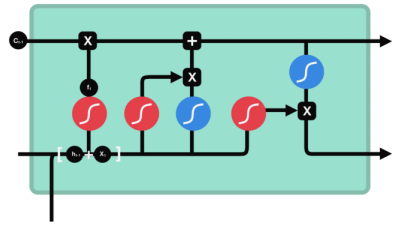


Deep Learned Process Parameterizations Provide Better Representations of Turbulent Heat Fluxes in Hydrologic Models

Andrew Bennett, Bart Nijssen

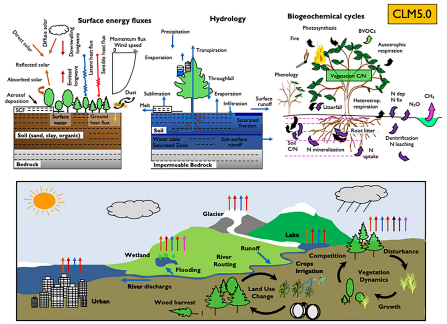
First published: 12 May 2021
https://doi.org/10.1029/2020WR029328

Use ML to represent one process in the process-based hydrologic model



Pure process-based models — Pure ML-AI models

It is not just black or white!

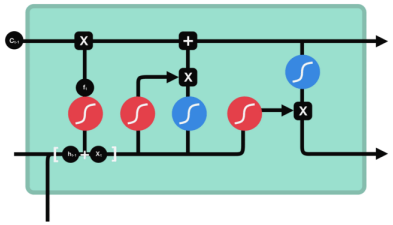


Deep Learned Process Parameterizations Provide Better Representations of Turbulent Heat Fluxes in Hydrologic Models

Andrew Bennett, Bart Nijssen

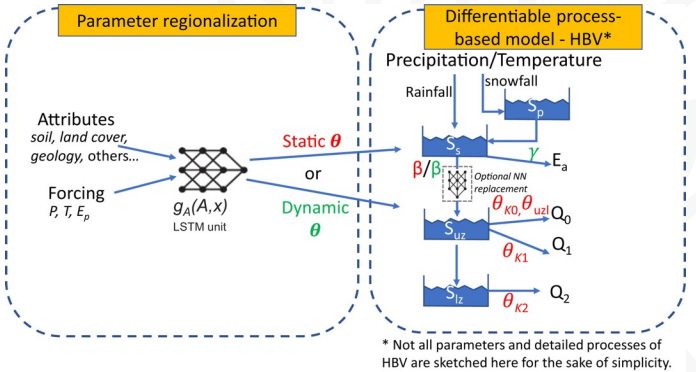
First published: 12 May 2021
https://doi.org/10.1029/2020WR029328

Use ML to represent one process in the process-based hydrologic model



Pure process-based models — Pure ML-AI models

Instead of having one big black box, multiple smaller black boxes are used to mimic the process of physically based hydrologic models



Which models do you prefer?