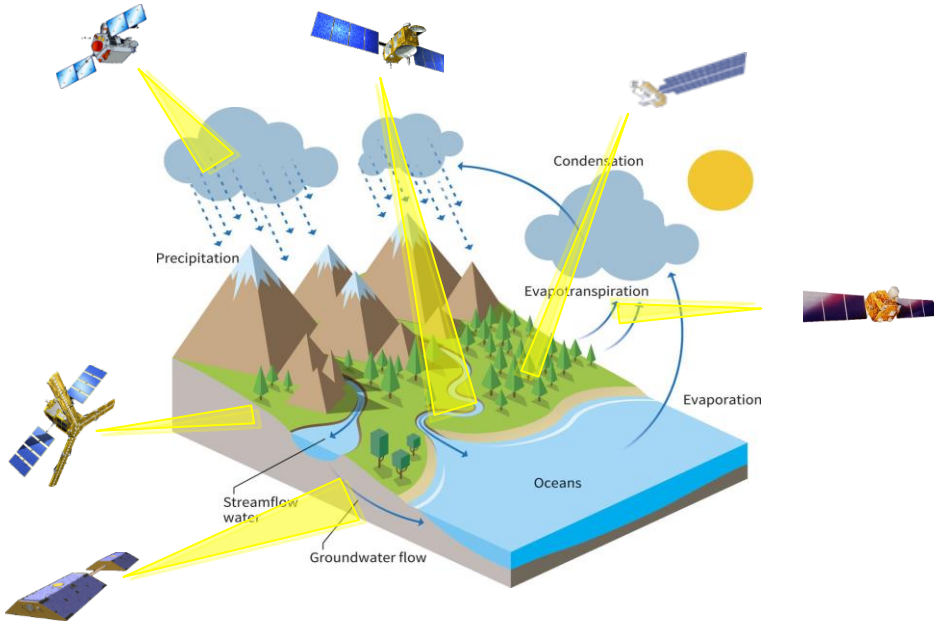
Improving remote sensing observations of the water cycle

with analytical methods, simple statistical models, and more complex machine-learning models



Matthew Heberger

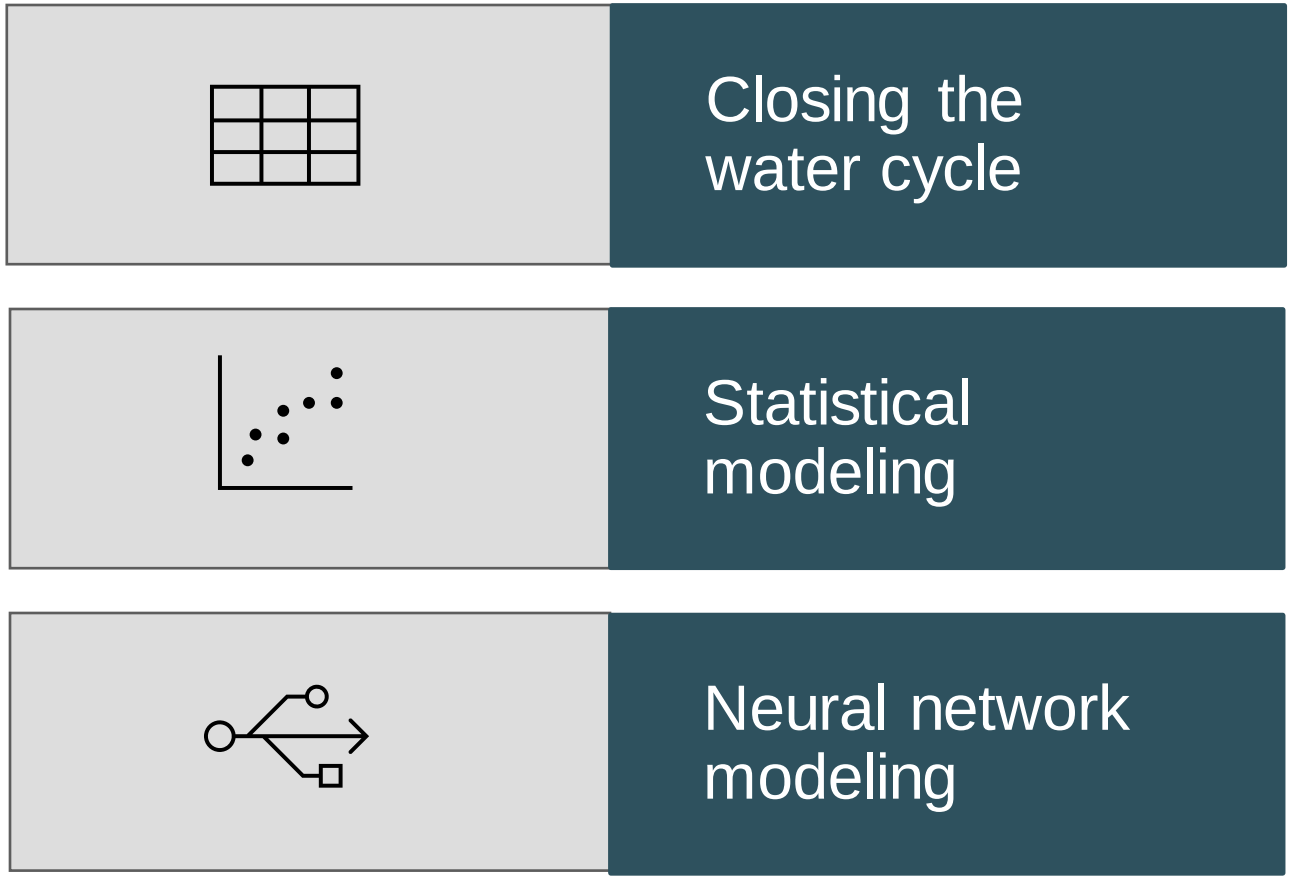
Filipe Aires

Victor Pellet

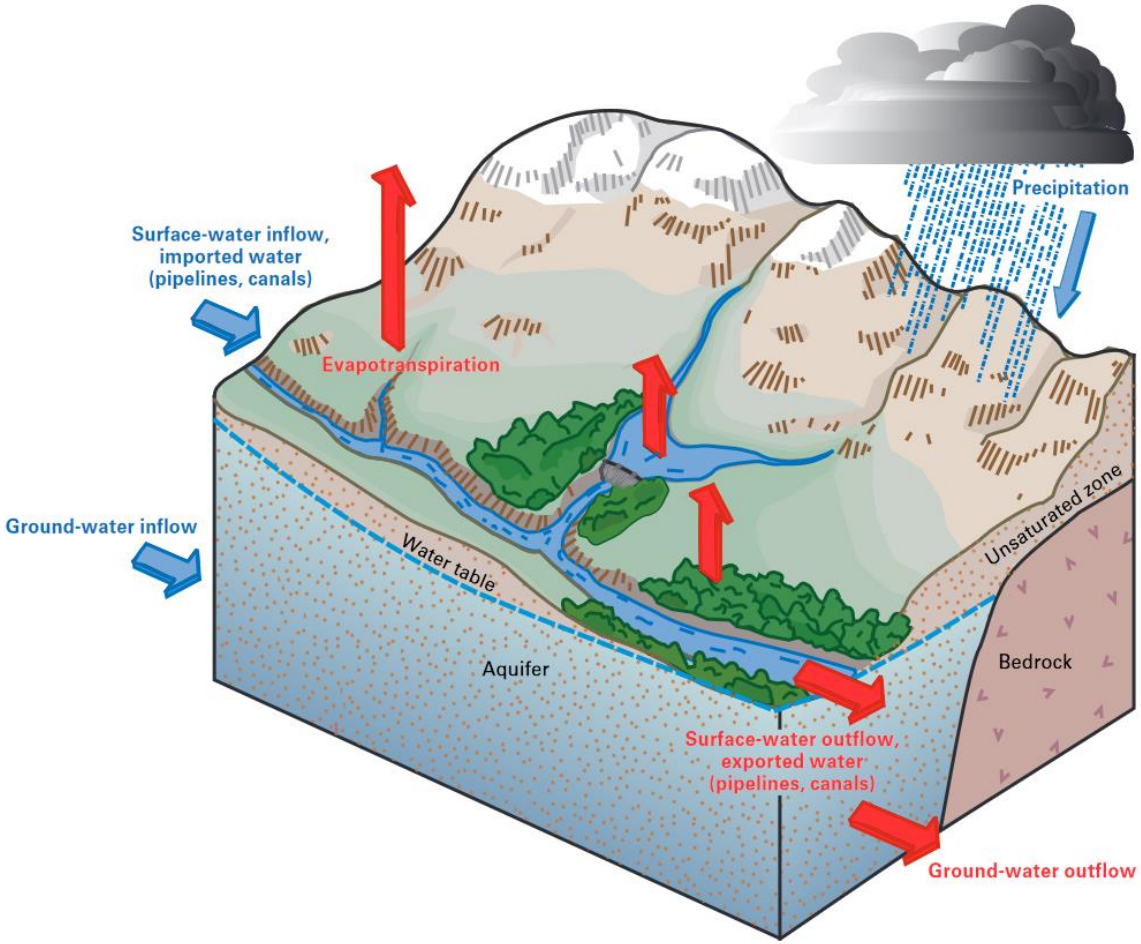
Paris Observatory & Sorbonne University

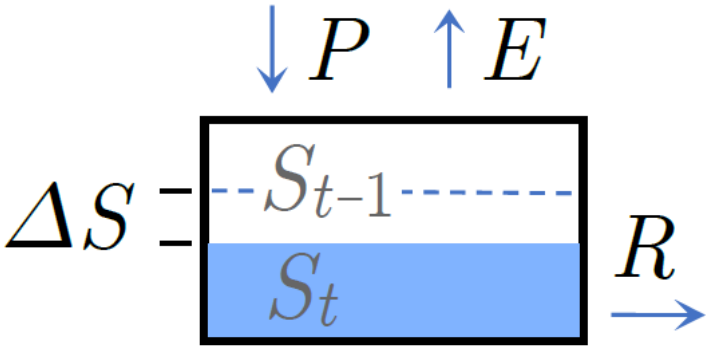
December 15, 2023

My PhD research focused on optimizing estimates of the water cycle globally, at the pixel scale



“Water budgets are important tools that water users and managers use to quantify the hydrologic cycle”*





$$P - E - \Delta S - R = 0$$

*Healy et al. (2007). Water Budgets: Foundations for Effective Water-Resources and Environmental Management. USGS.

The fundamental problem: remote sensing datasets are "incoherent" – meaning the water cycle is not closed

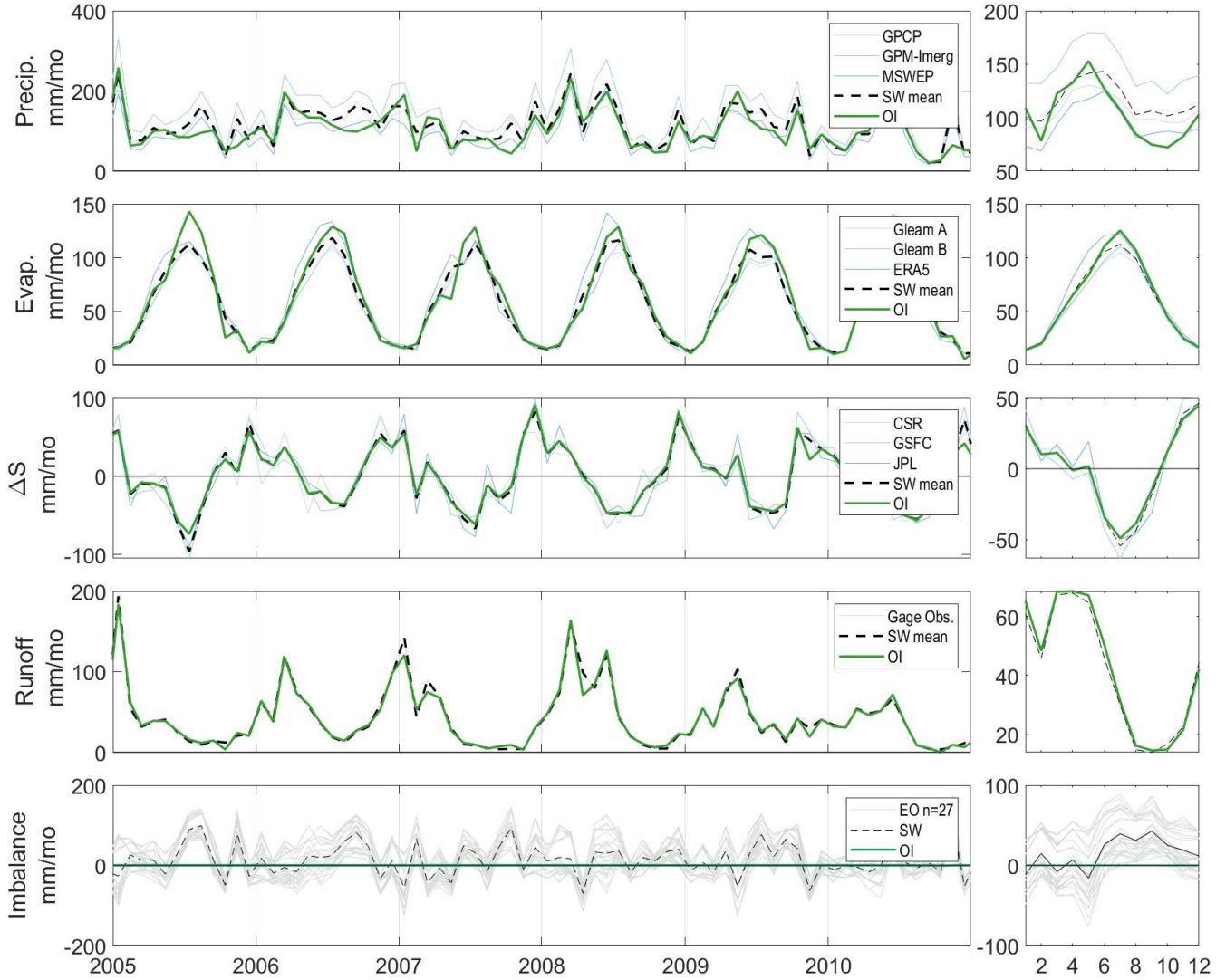
Precipitation, P

Evapotranspiration, E

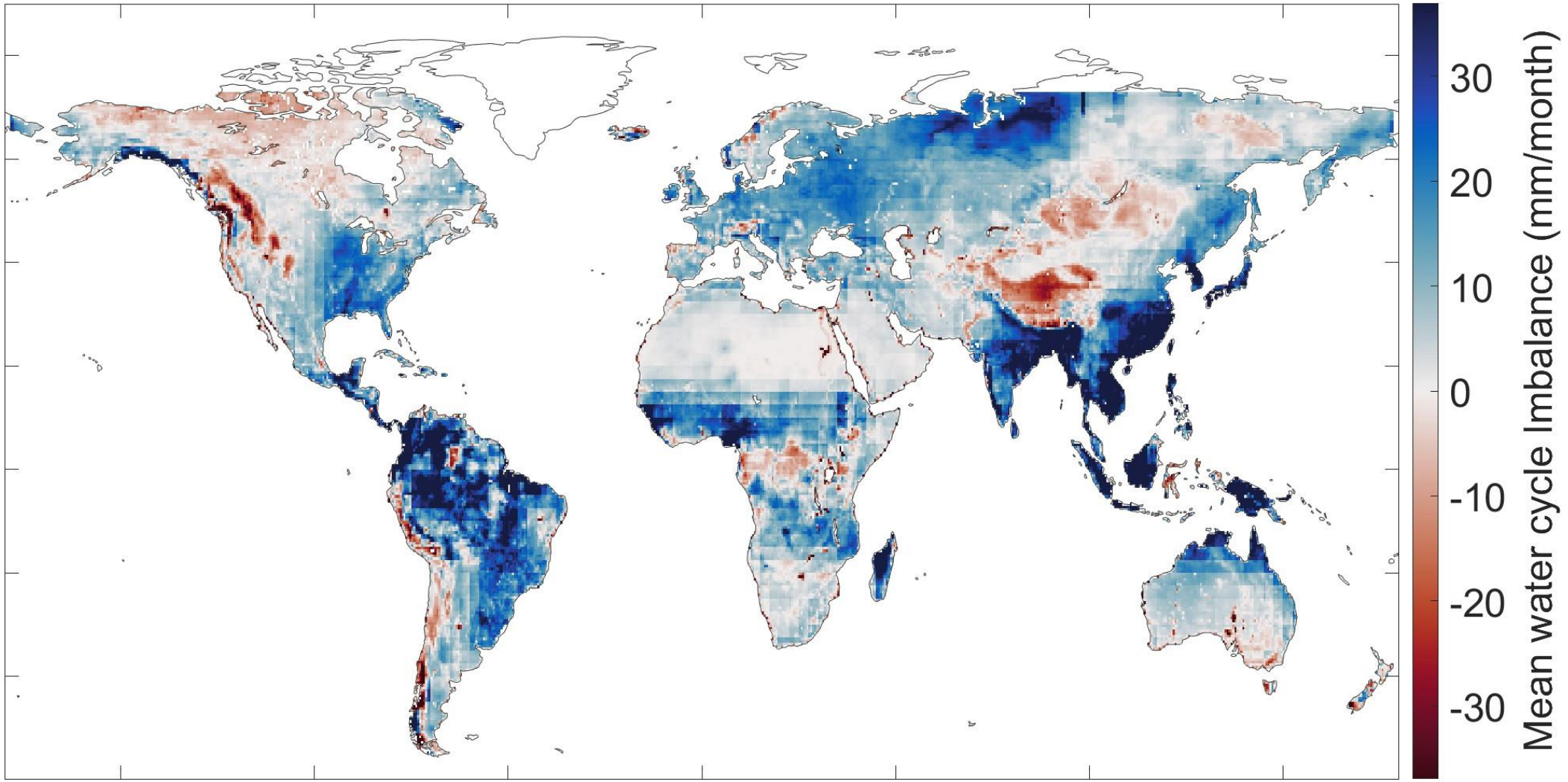
Total Water Storage Change, ΔS

Runoff, R

Imbalance, I



We see strong spatial patterns in the water cycle “imbalance”



Researchers have used a variety of methods to solve this problem, but to date, no consensus has emerged on which is best

"Do nothing"

Best combination

Bias correction

Ensemble-based methods

Include energy budget constraints

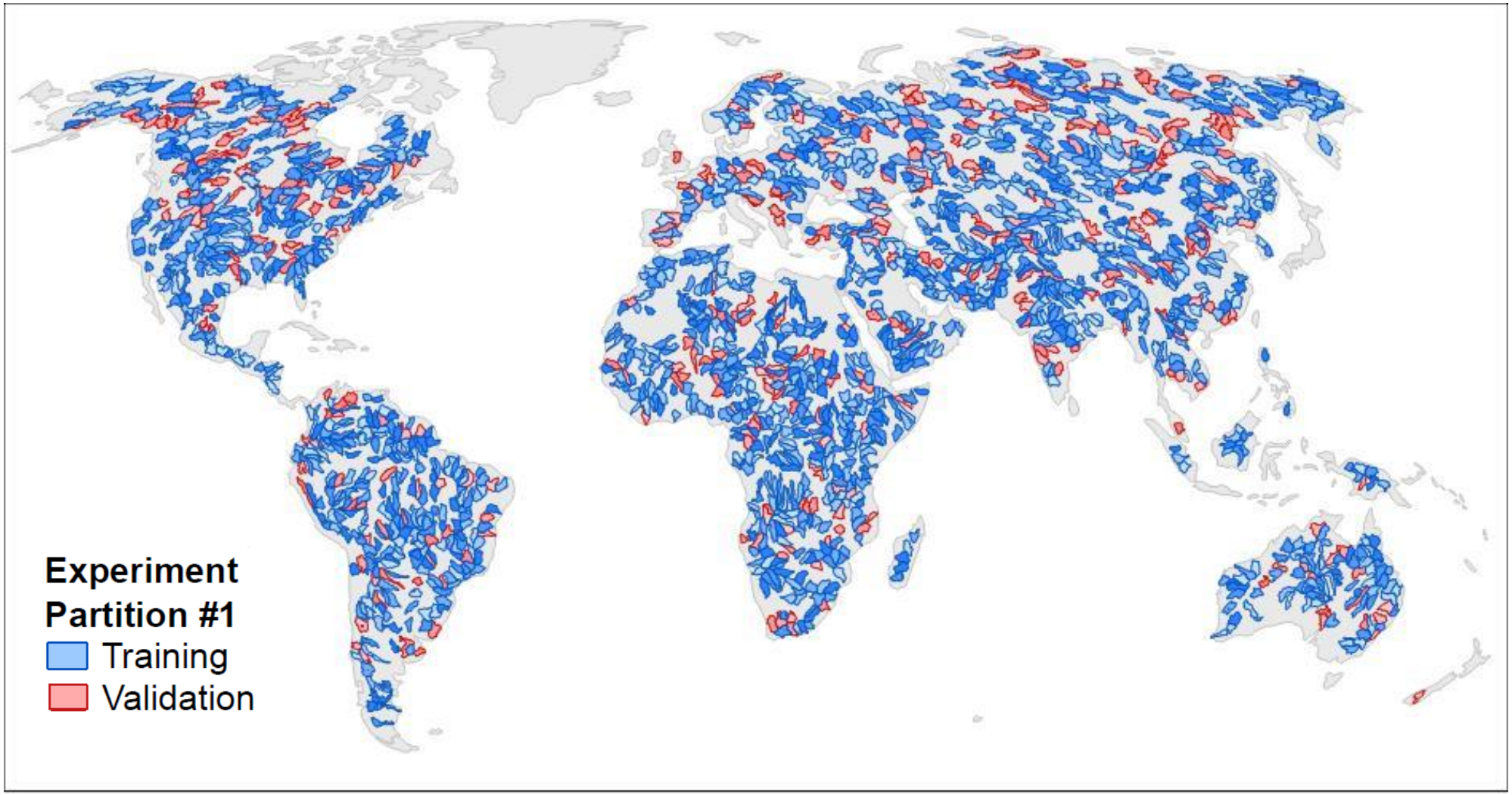
Statistical optimization

Assimilation modeling

This study's input datasets include several gridded datasets from remote sensing and models, covering 2000 to 2019

Data Set	Begin	End	Temporal resolution	Spatial resolution	Reference
Precipitation					
GPCP v2.3	1979	present	daily, monthly	2.5°	Adler et al. 2018.
GPM-IMERG	2000	present	daily	0.10º	Huffman et al. 2019
MSWEP	1979	present	daily, monthly	0.10°	Beck et al. 2019.
Evapotranspiration					
GLEAM v3.5A	1980	present	daily	0.25º	Miralles et al. 2011; Martens et al. 2017
GLEAM v3.5B	2003	present	daily	0.25º	idem.
ERA5	1950	present	3-hour, daily, monthly	0.25º	Hersbach, et al. 2018.
Water Storage					
GRACE-CSR	2002	present	quasi-monthly*	0.25°	Save, Bettadpur, and Tapley, 2016
GRACE-JPL	2002	present	quasi-monthly*	0.50°	Landerer, 2021; Landerer et al. 2020
GRACE-GSFC	2002	present	quasi-monthly*	0.50º	Loomis, Luthcke, and Sabaka, 2019
Runoff					
GRUN	1902	2019	monthly	0.5°	Ghiggi et al. 2021

This study was conducted over 1,698 river basins, ranging in size from 20,000 to 50,000 km²



Optimal interpolation is a closed-form analytical solution that modifies water cycle components to close the water budget

$$\mathbf{X} = \begin{bmatrix} P \\ E \\ \Delta S \\ R \end{bmatrix} \qquad \mathbf{G} = \begin{bmatrix} 0 \\ -1 \\ -1 \\ -1 \end{bmatrix}$$

one set of observations,
(one basin, one month)

$$\mathbf{X}^T \mathbf{G} = 0$$

$$P - E - \Delta S - R = 0$$

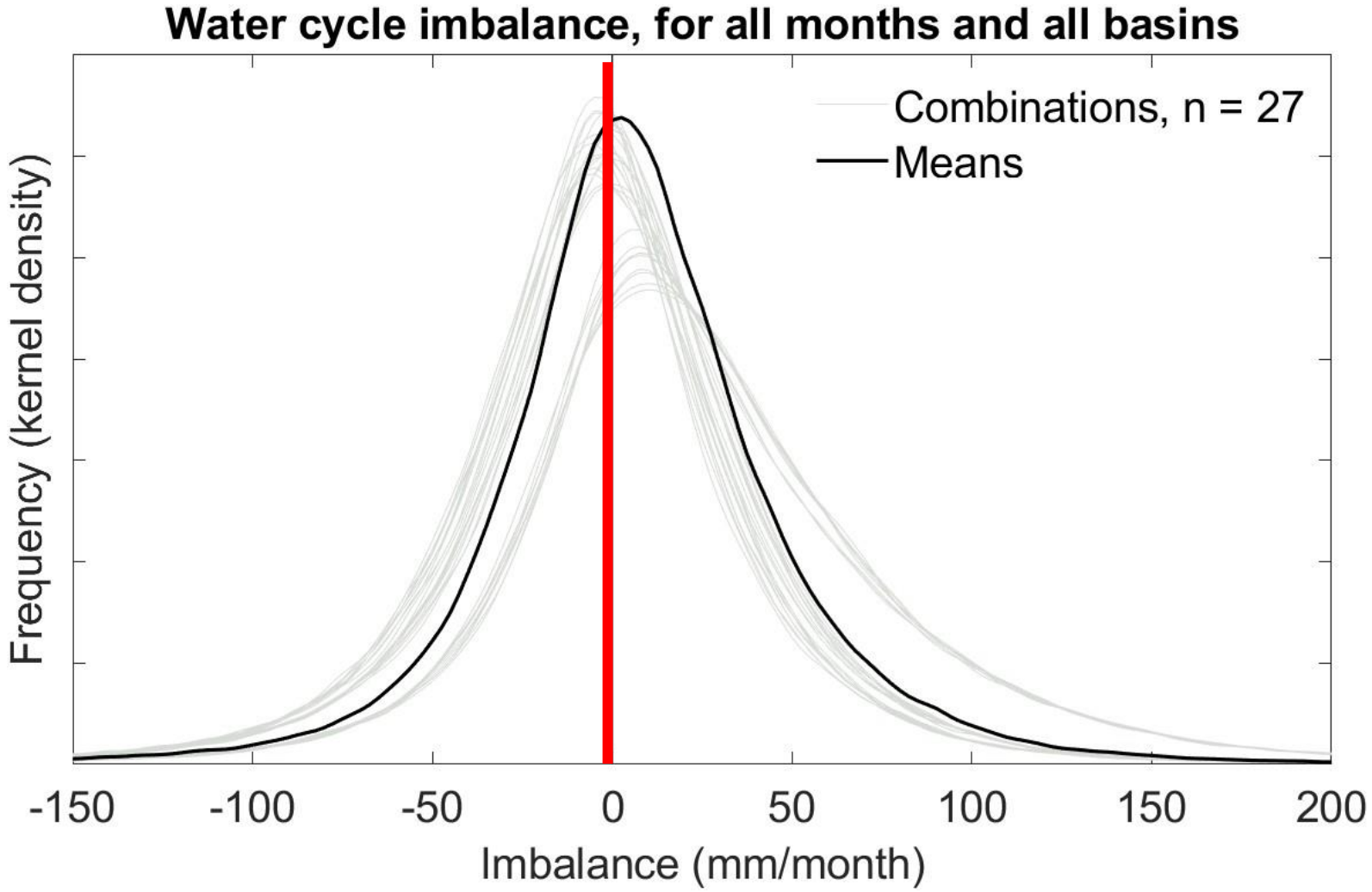
$$\mathbf{X}_{OI} = \mathbf{K}_{PF} \cdot \mathbf{X}$$

$$\mathbf{K}_{PF} = \mathbf{I} - \mathbf{B}\mathbf{G}^T \left(\mathbf{G}\mathbf{B}\mathbf{G}^T\right)^{-1} \mathbf{G}$$

$$\mathbf{B} = \begin{bmatrix} \sigma_P^2 & 0 & 0 & 0 \\ 0 & \sigma_E^2 & 0 & 0 \\ 0 & 0 & \sigma_{\Delta S}^2 & 0 \\ 0 & 0 & 0 & \sigma_R^2 \end{bmatrix}$$

Post-filter matrix has two components:
(1) Drive water cycle residual to zero
(2) Make the minimum changes necessary to the water cycle components

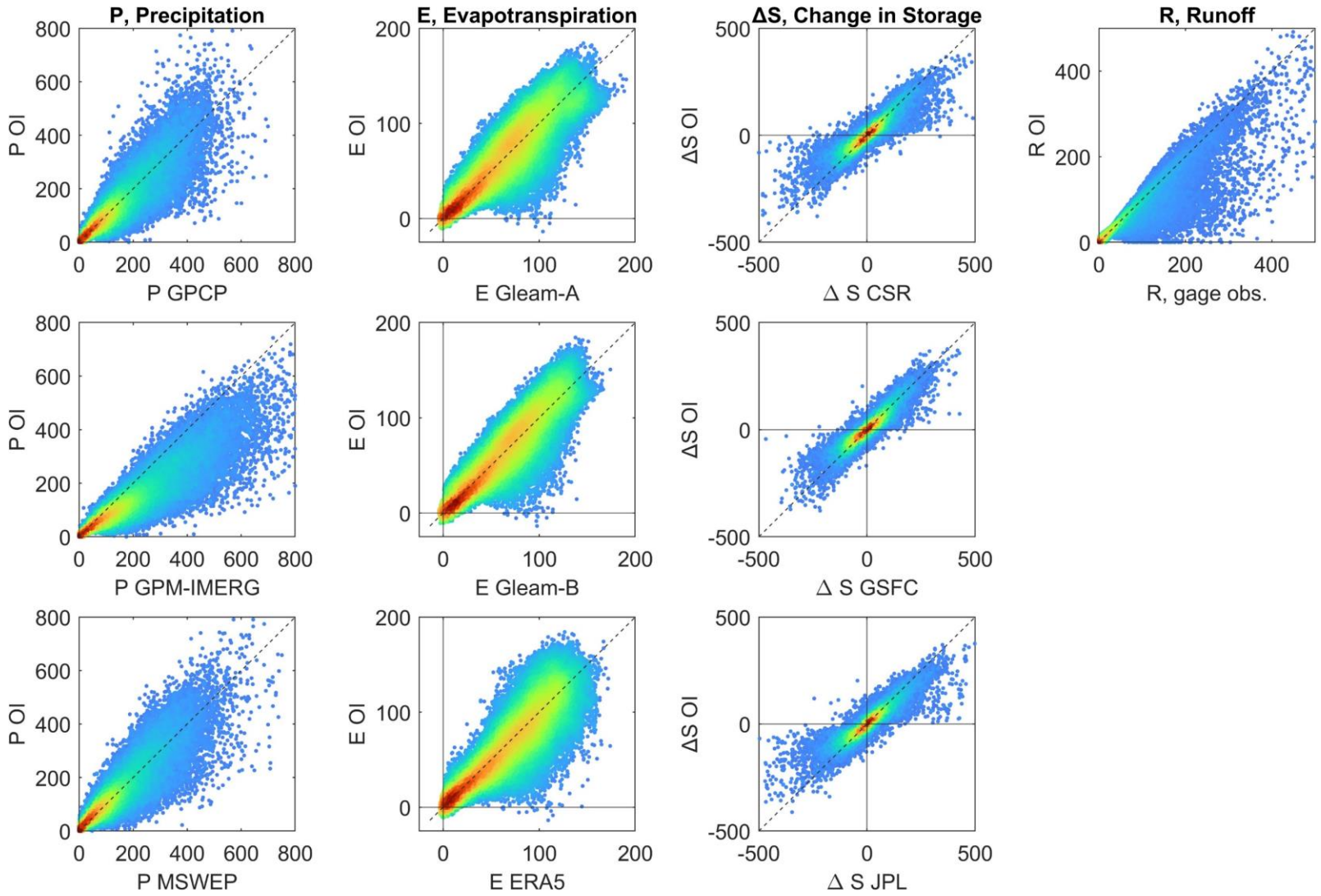
Optimal interpolation does an excellent job at closing the water cycle *at the basin scale*



Before:
$I = 6 \pm 47$ mm/month

After
$I = 0$ for all I

…usually without changing the original observations too much



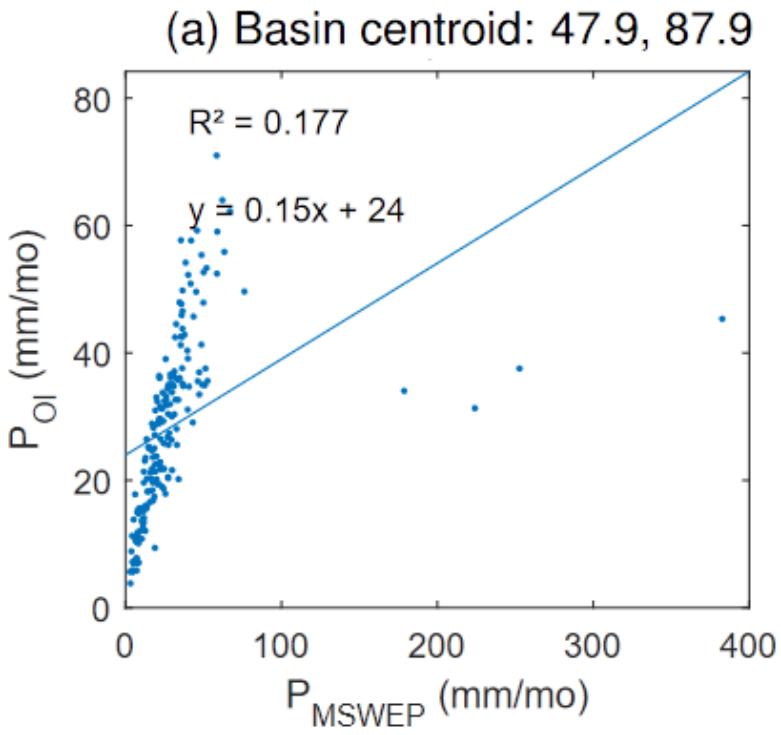
Optimal interpolation is appealing because

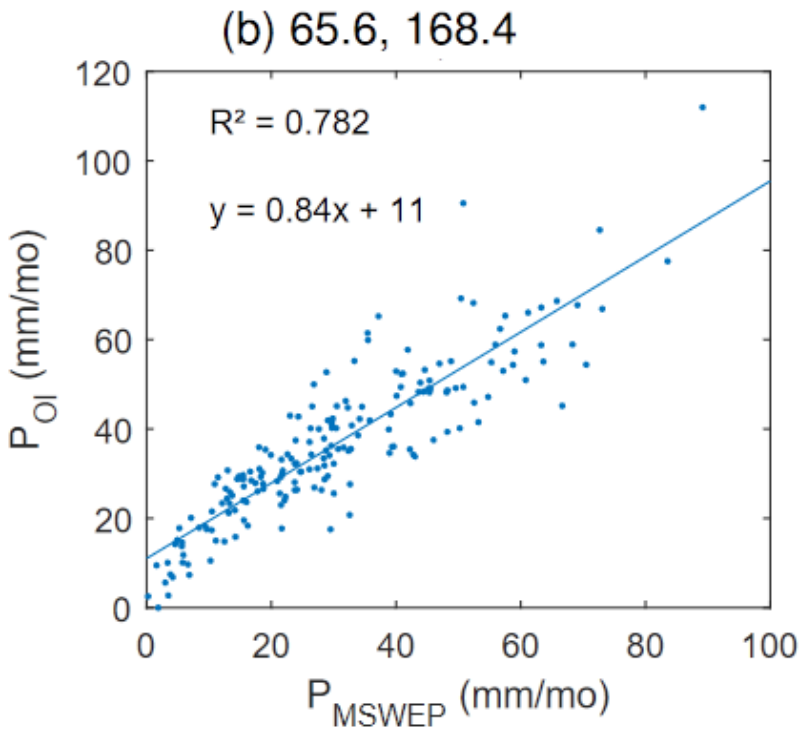
- It’s simple
- It has a basis in information theory
- It exploits information on uncertainties in each water cycle component
- It makes the smallest changes necessary to achieve closure

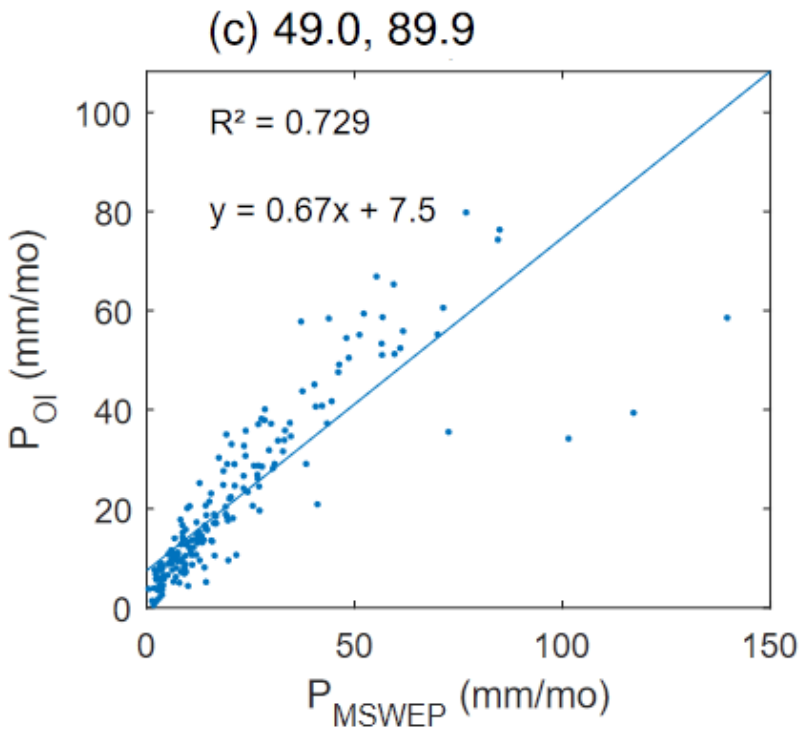
But!

- It requires all 4 water cycle components ☹
- This means it can only be applied over river basins, where we have observations of river discharge.

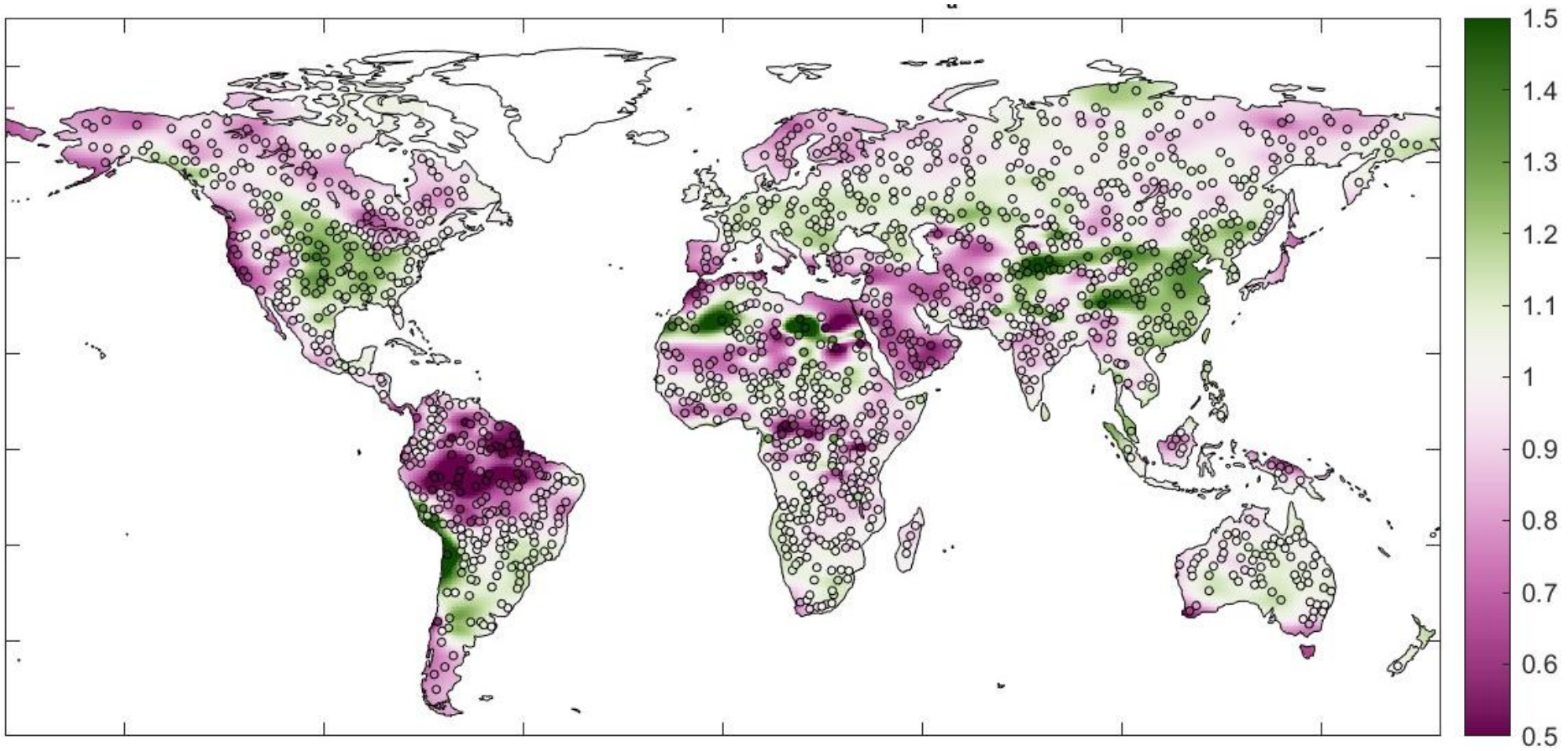
To extrapolate this solution to the pixel scale, we tried using simple linear models plus spatial interpolation





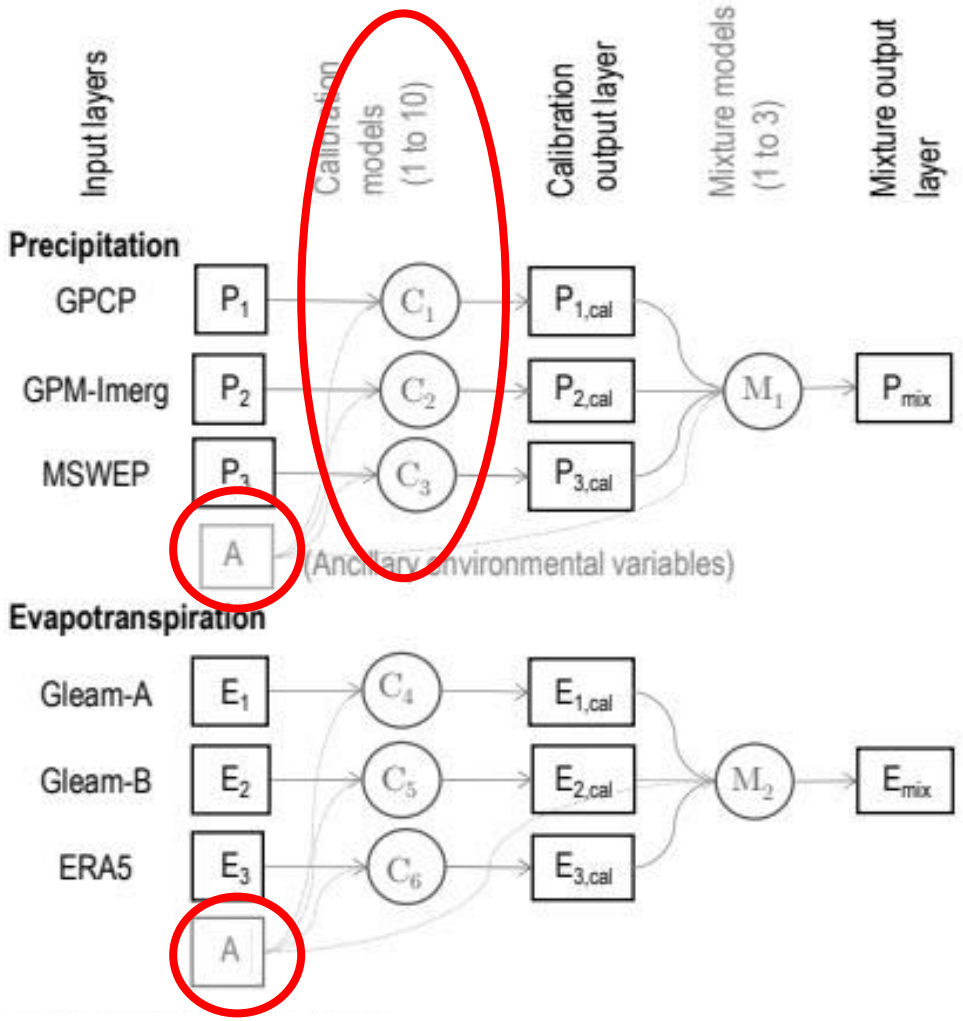


The regression parameters for each variable were estimated at the pixel scale with spatial interpolation



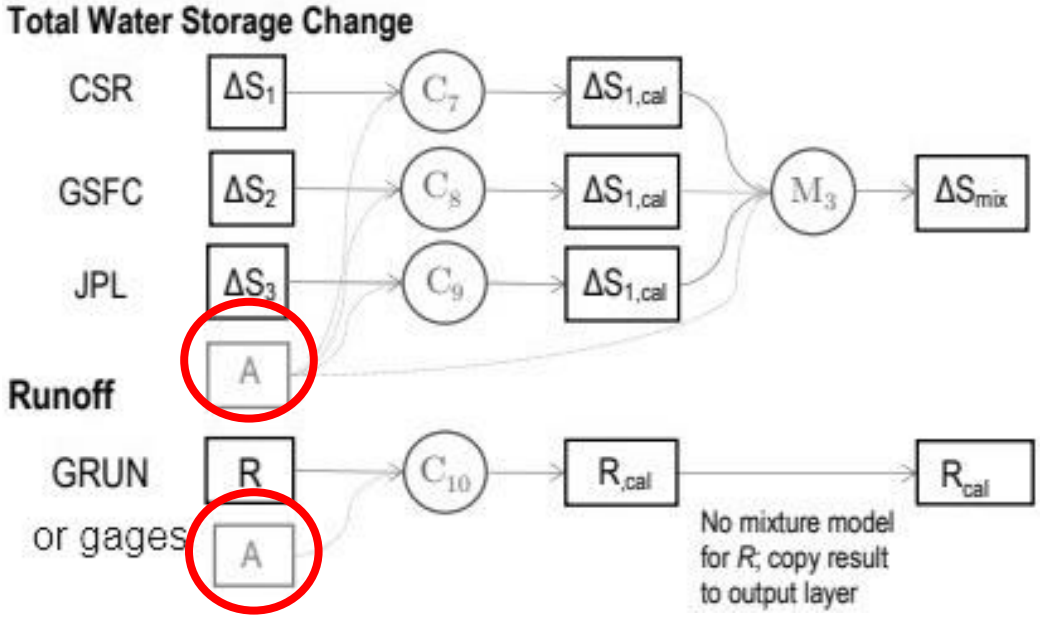
Interpolated slope for GLEAM A Evapotranspiration

We fit neural network models to remote sensing datasets to approximate the solution from optimal interpolation



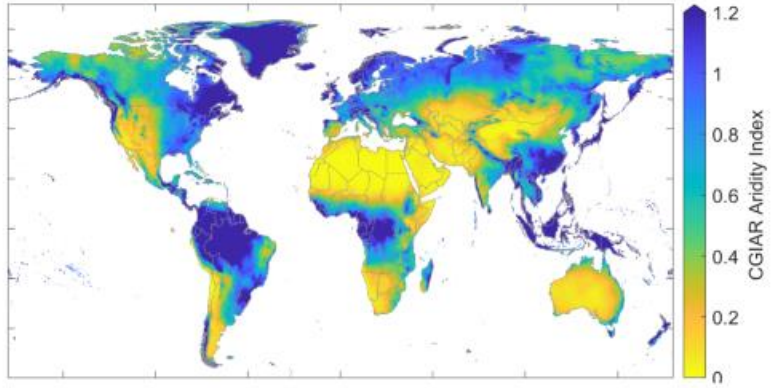
There are separate layers for calibration and mixture

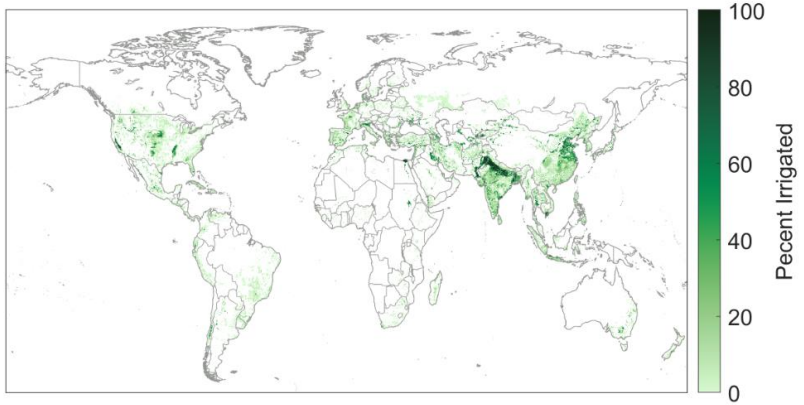
We used ancillary environmental variables to describe the local environment and improve the model's fit



Ancillary environmental data includes static and time-varying variables that have a clear link to the hydrologic cycle

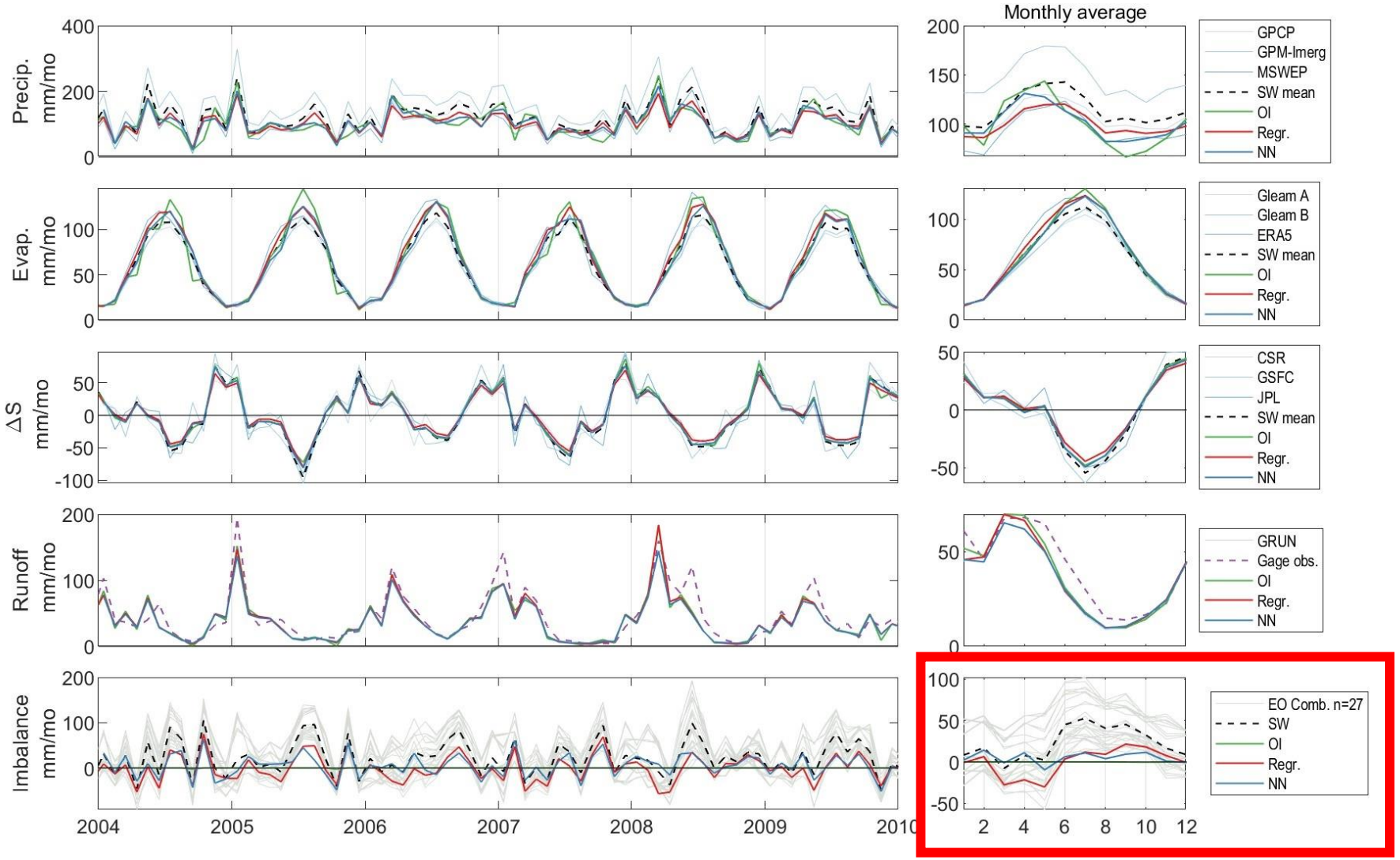
#	Variable	Units	Source
1	Aridity index	dimensionless	calculated
2	Elevation, basin mean	meters	Amatulli et al. (2018)
3	Latitude, basin centroid	decimal degrees	calculated
4	Slope, basin median	dimensionless	Amatulli et al. (2018)
5	Vegetation Index, EVI	dimensionless	Didan (2015)
6	Irrigated area (percent)	dimensionless	Siebert et al. (2015)
7	Longitude, basin centroid	decimal degrees	calcualted
8	Burned area (percent)	dimensionless	Giglio et al. (2020)
9	Snow cover (percent)	dimensionless	Hall and Riggs (2021)
10	Solar radiation	J/m²	Hogan (2015)
11	Temperature	℃	Wan et al. (2021)
12	Vegetation growth/senescence	dimensionless	calculated



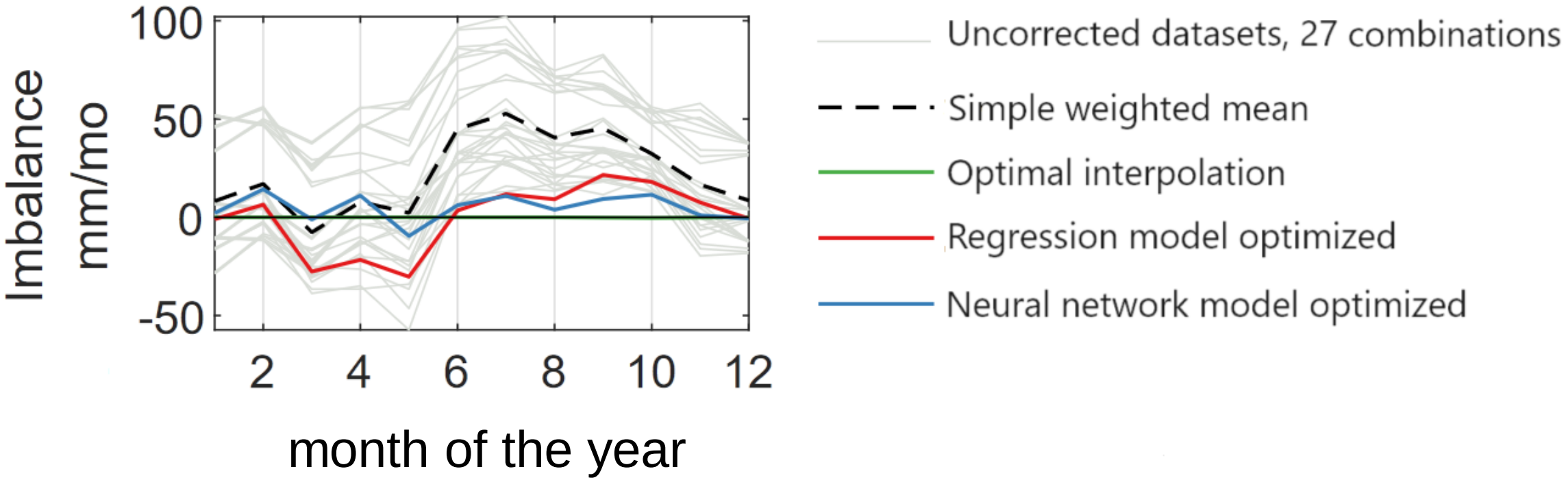


8 out of the 12 variables significantly improved the fit of the neural network model

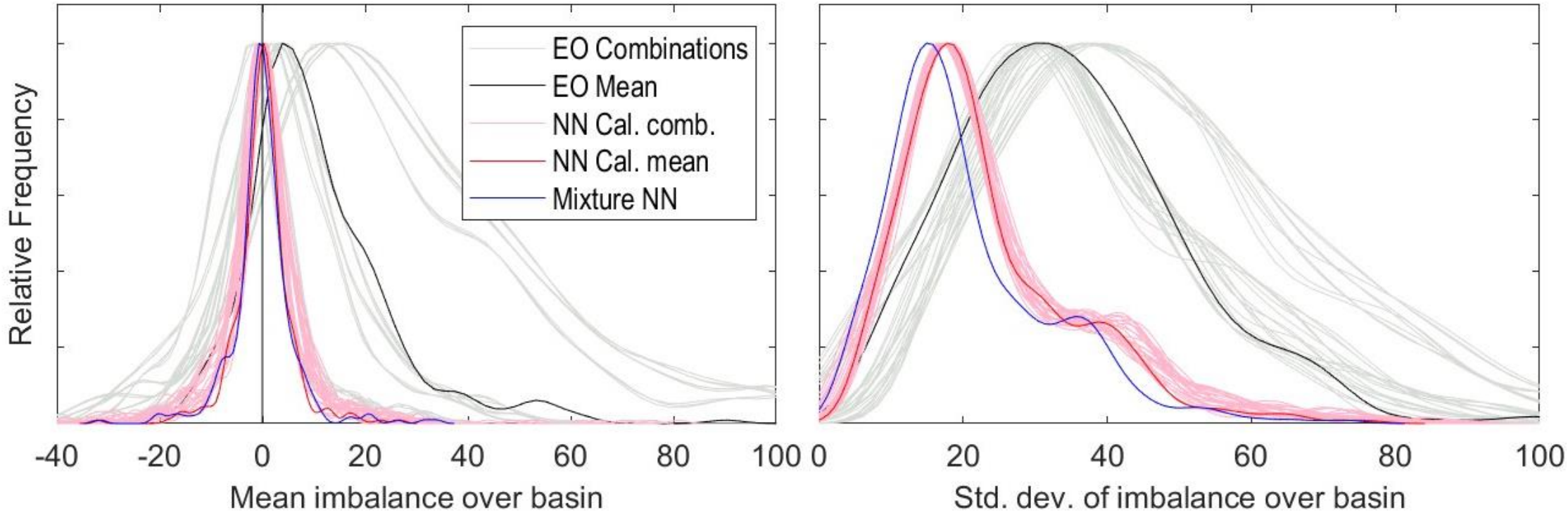
Here is an example of the output, over a single river basin, the White River at Petersburg, Indiana, USA (29,000 km²)



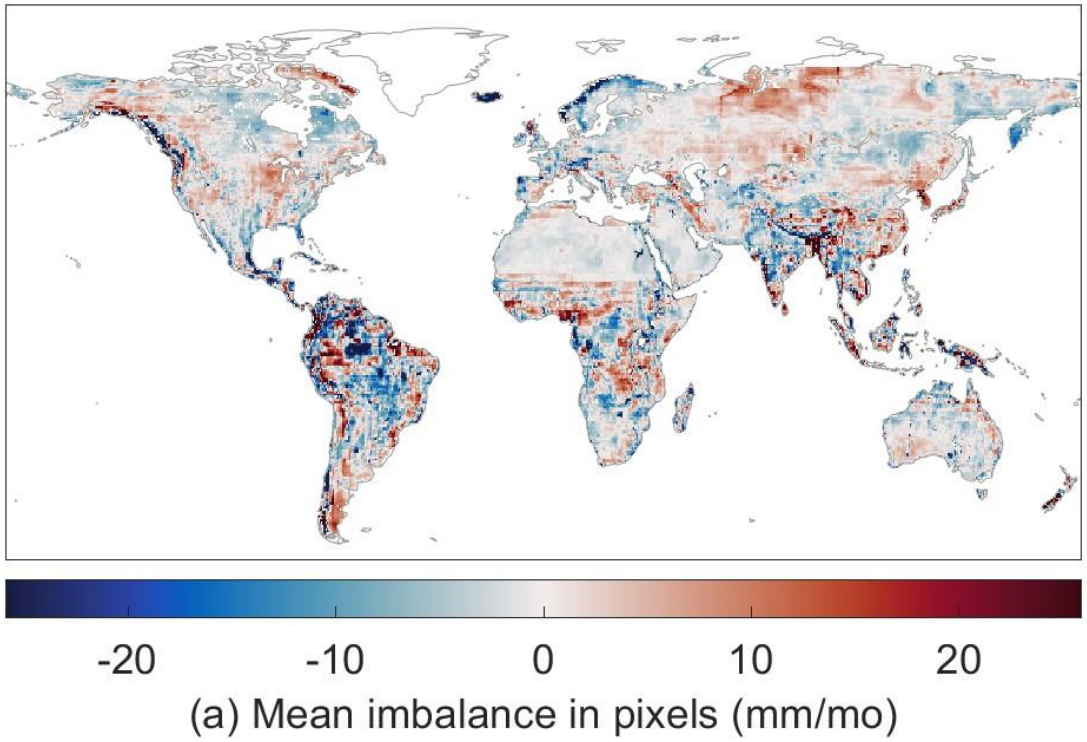
Here is an example of the output, over a single river basin, the White River at Petersburg, Indiana, USA (29,000 km²)



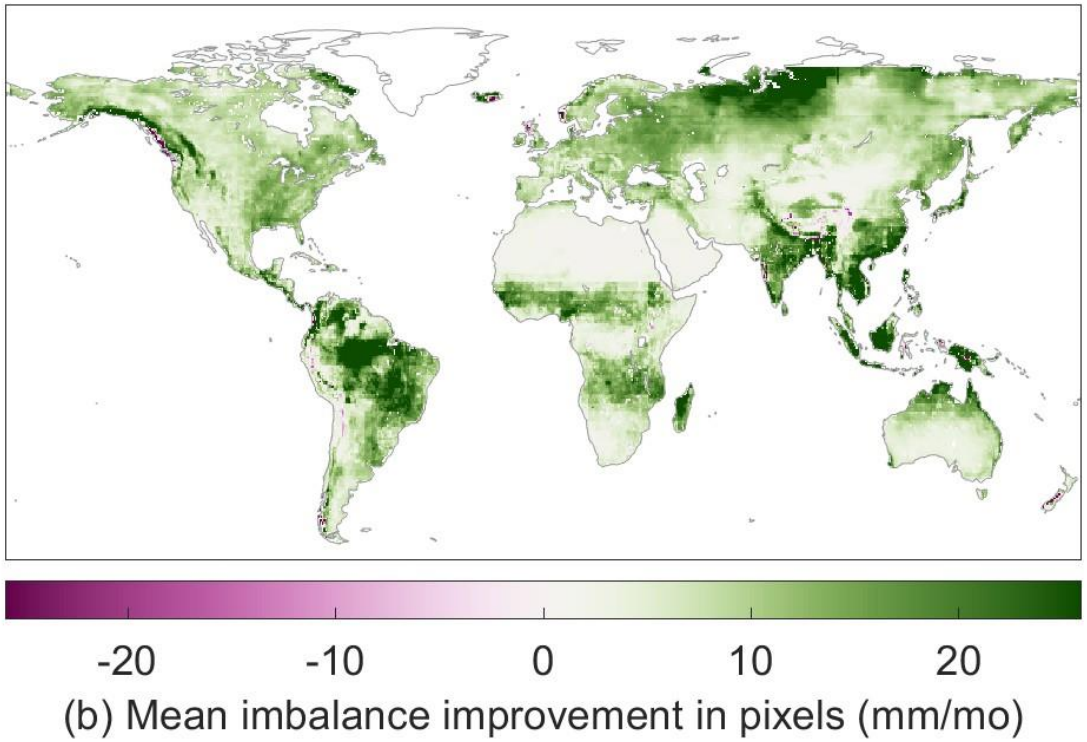
At the river basin scale, our model reduces the mean and variance of the water cycle “imbalance”



At the pixel scale, the imbalance in the water cycle is improved *almost* everywhere



(a) Mean imbalance in pixels (mm/mo)

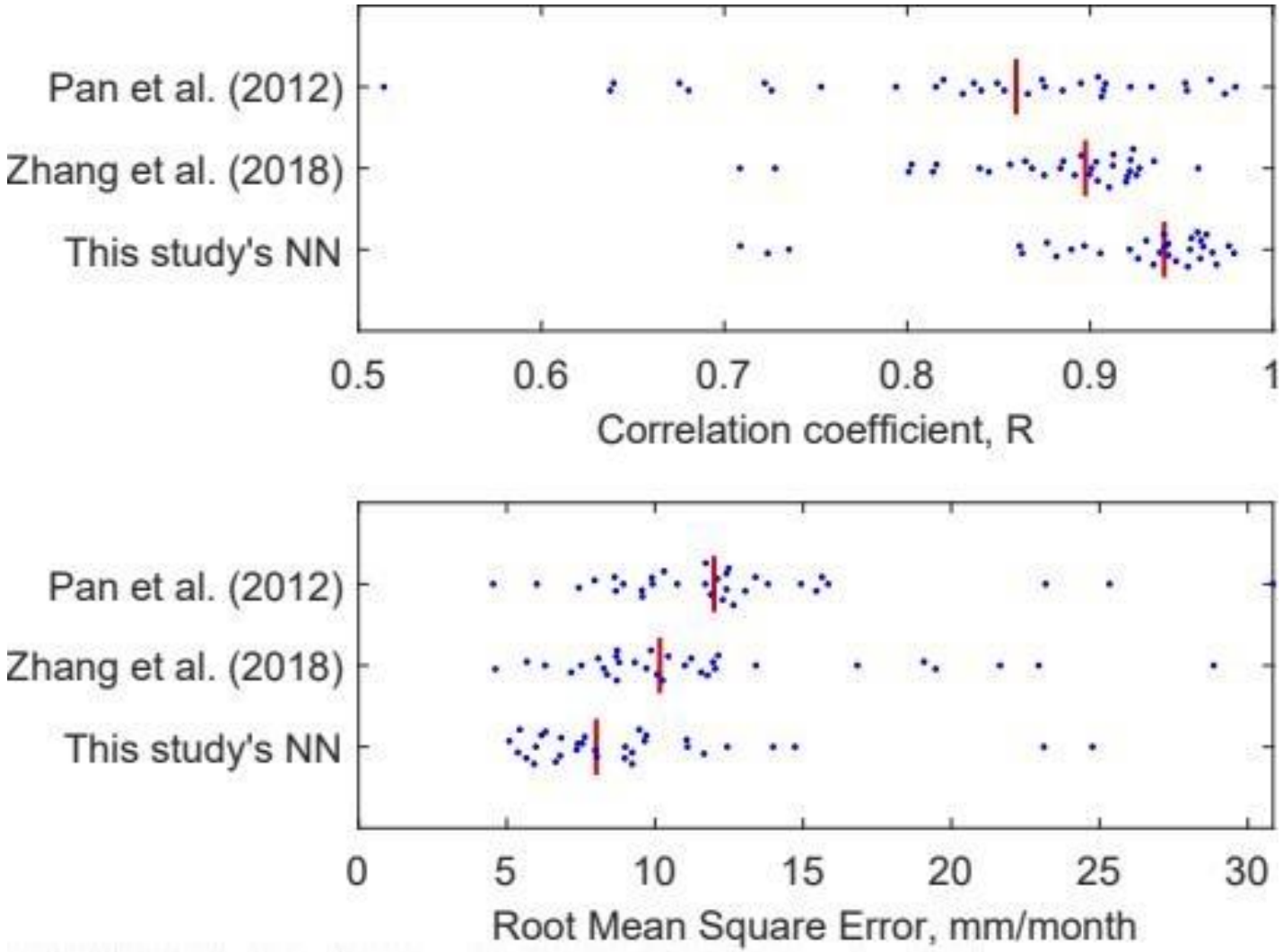


(b) Mean imbalance improvement in pixels (mm/mo)

We also evaluated our recalibrated EO datasets by comparing them to ground-based observations of *P*, *E*, and *R*.

Dataset	NSE	RMSE, mm/mo	Percent Bias
Evapotranspiration, at *n* = 117 flux towers			
GLEAM-A	0.65	21.4	3.6%
GLEAM-A (Regr. cal)	**0.70**	19.2	7.8%
GLEAM-A (NN cal)	0.69	**19.0**	**3.3%**
GLEAM-B	0.69	20.1	5.4%
GLEAM-B (Regr. cal)	0.67	19.9	**4.9%**
GLEAM-B (NN cal)	0.69	**18.5**	6.1%
ERA5	0.70	19.9	7.6%
ERA5 (Regr. cal)	0.68	20.9	7.8%
ERA5 (NN cal)	0.70	**19.4**	**6.0%**
EO SW mean	0.70	19.5	**3.9%**
Regr. cal. avg.	0.70	20.1	4.9%
NN Mixture Model	0.69	**19.4**	4.1%

For TWSC, our model works as well as, and sometimes better, than state of the art assimilation models

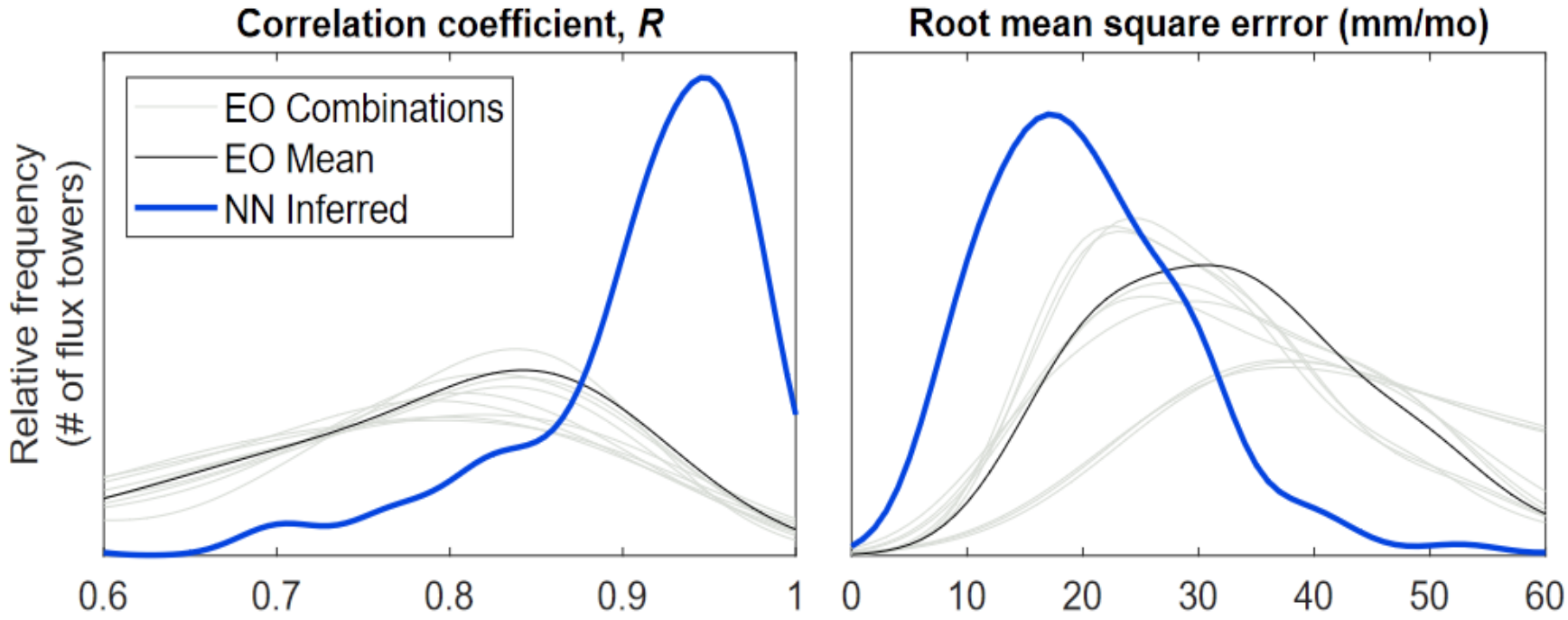


The neural network model gives a higher correlation *R*, and lower root mean square error over 32 large river basins

Water-budget based methods can be used to estimate missing water cycle components

- For example, we can calculate basin evapotranspiration with $E = P - \Delta S - R$
- We found that such estimates are significantly improved when using the neural network-calibrated datasets, compared to using uncorrected remote sensing data.

Goodness of fit of *E* estimated by the water budget method, compared to observed *E* at 117 flux towers



In summary, statistical and machine learning models can help “close the water cycle” at the global scale

Statistical models allow us to “recalibrate” and optimize remote sensing data at both the river basin and pixel scale.

The results can be used to create water budgets, estimate missing water cycle components, or to show where satellite datasets are biased and could potentially be improved.

For data, code, my thesis, and contact info:
https://mghydro.com

Questions?