Journal Club
01. June 2021

machine learning in climate science

Markus Deppner

Today's paper

Monthly streamflow forecasting using Gaussian Process Regression

Alexander Y. Sun [a,*], Dingbao Wang [b], Xianli Xu [c,d]

[a] Bureau of Economic Geology, Jackson School of Geosciences, University of Texas Austin, Austin, TX 78713, United States
[b] Department of Civil, Environmental, and Construction Engineering, University of Central Florida, Orlando, FL 32816, United States
[c] Key Laboratory for Agro-Ecological Processes in Subtropical Region, Institute of Subtropical, Agriculture, Chinese Academy of Sciences, Changsha, China
[d] Huanjiang Observation and Research Station for Karst Ecosystem, Chinese Academy of Sciences, Guangxi, China

Motivation for this paper

- Streamflow forecasting essential in water management and resource planning
- Gaussian Process Regression compared to Linear Regression and artificial neural networks (ANN)
- Loss of predictability in recent years due to a changing climate and anthropogenic activities.
- A major challenge of streamflow prediction stems from the fact that streamflow is a temporally lagged, spatial integral of runoff over a river basin
- Demonstrate efficacy of GPRs
- Analyse factors that can potentially affect basin streamflow predictability

Existing approaches and methods

- Physics based methods
  - Mathematical abstractions of physical processes that determine water movement and storage in watersheds
- Time series methods
  - Linear Regression models (short-term forecasting - daily, weekly)
  - Cannot handle nonlinearity by rainfall-runoff models
- Machine learning methods
  - Data-driven
  - ANN - tendency to overfit and unstable for short training data records

Motivation to use Gaussian Process Regression

- Usually deterministic algorithms which do not provide quantification of uncertainty.
- Use of ARMA models, Kalman filters, RBF networks in the past, which can be seen as a sequential version of GP-based learning algorithms.
- GP provides three in one - hyperparameter estimation, model training, uncertainty estimation
- Demonstrate efficacy of GPRs

Data

- MOPEX database
  - 438 basins across the U.S.
  - from 01. Jan. 1948 until 31.Dec. 2003
  - basin averaged daily hydrometeorological data (streamflow, precipitation, min- & max temperature, potential evaporation)
- Extensions of MOPEX from Jan 2004 until Dec. 2012

General Regression task

- Regular Regression task :

  Learn input output mapping of d-dimensional predictor $\mathbf{x} \in \mathbb{R}^d$ and target variable y

$$y = f(\mathbf{x}),$$

- General function f as a linear combination of basis functions (linear or non-linear) and scaling weights for each basis function

$$\hat{f}(\mathbf{x}, \mathbf{w}) = \sum_{j=1}^{M} w_j \phi_j(\mathbf{x}),$$

- Additional error term

$$y = \sum_{j=1}^{M} w_j \phi_j(\mathbf{x}) + \varepsilon,$$

Gaussian Process Regression

- Model outputs corresponding to the input dataset X $\mathbf{f} = \{\hat{f}(\mathbf{x}_i, \mathbf{w})\}_{i=1}^{N}$

$$\hat{f}(\mathbf{x}_i, \mathbf{w}) = \sum_{j=1}^{M} w_j \phi_j(\mathbf{x}_i), \quad i = 1, \ldots, N$$

- Lets define a N x M design matrix $\mathbf{\Phi}$ that contains the output of the basis functions for respective input

$$\mathbf{f} = \mathbf{\Phi}\mathbf{w} \qquad \phi_j = [\phi_1(\mathbf{x}_i), \phi_2(\mathbf{x}_i), \ldots, \phi_M(\mathbf{x}_i)], \quad j = 1, \ldots, N$$

- GPR can be described by it second-order statistics with mean $m(\mathbf{x})$ and $k(\mathbf{x}, \mathbf{x}')$ as the covariance function $f(\mathbf{x}) \sim \mathrm{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$
- Any finite subset of a GP has joint Gaussian distribution.
- Then the prior distribution of f is Gaussian

$$p(\mathbf{f}|\mathbf{X}, \theta) \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$$

Gaussian Process Regression

- Then the prior distribution of f is Gaussian

θ as the hyperparameters for covariance function

$$p(\mathbf{f}|\mathbf{X},\theta) \sim \mathcal{N}(\mathbf{0},\mathbf{K})$$

- Mean of zero is chosen here for convenience - in general any mean function possible
- Covariance Matrix can be written as inner product with respect to $\mathbf{\Sigma}_w$ which is the covariance matrix of w

$$\mathbf{K} = \mathbf{\Phi} E(\mathbf{w}\mathbf{w}^T)\mathbf{\Phi}^T = \mathbf{\Phi}\mathbf{\Sigma}_w\mathbf{\Phi}^T$$

Gaussian Process Regression

- If the model error is independent and identically Gaussian distributed than y becomes Gaussian

$$p(\mathbf{y}|\mathbf{f}, \sigma^2) \sim \mathcal{N}(\mathbf{f}, \sigma^2\mathbf{I})$$

$\boldsymbol{\sigma}^2$ variance of model error

- Desired posterior distribution of our GP is

$$p(\mathbf{f}|\mathbf{y}, \mathbf{X}, \theta, \boldsymbol{\sigma}^2) = \frac{p(\mathbf{y}|\mathbf{f}, \sigma^2)p(\mathbf{f}|\mathbf{X}, \theta)}{p(\mathbf{y}|\mathbf{X}, \theta, \sigma^2)}$$

- Prior and likelihood are Gaussian which is why posterior distribution is Gaussian as well.

Gaussian Process Regression

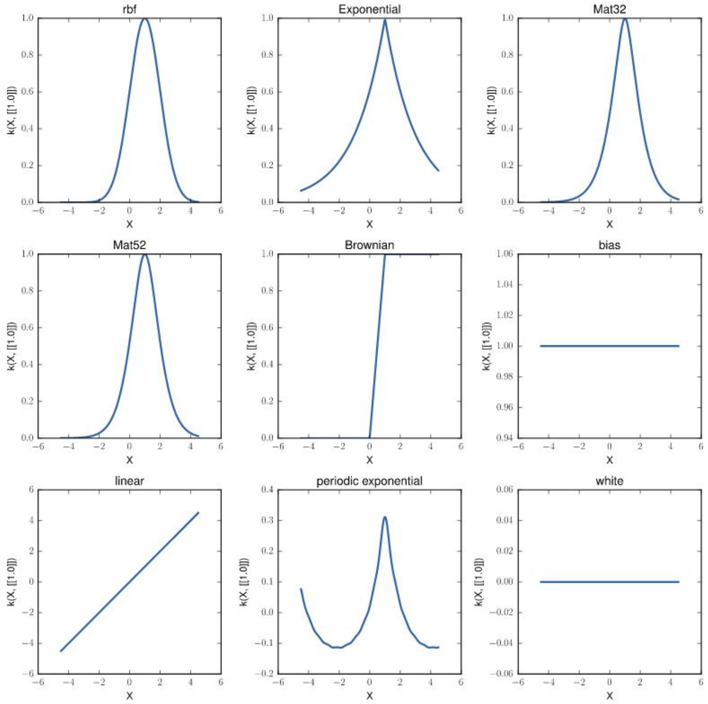
- Mean and Covariance are obtained by substituting prior and likelihood into Bayes' rule

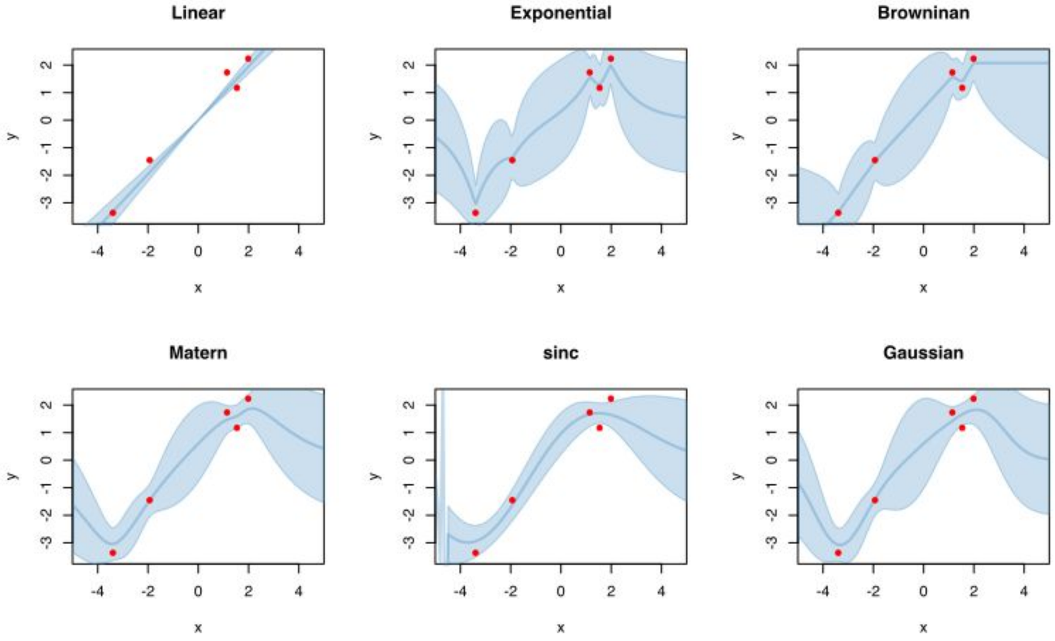
$$\mu = \mathbf{K}^T (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{y}$$

$$\boldsymbol{\Sigma} = \mathbf{K} - \mathbf{K}^T (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{K}$$

- Shift from determining the basis functions and weights to determining the covariance matrix K.

Gaussian Process Regression





https://nbviewer.jupyter.org/github/SheffieldML/

http://sashagusev.github.io/2016-01/GP.html

Gaussian Process Regression

- The missing marginal probability can be computed by integration over f

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{f}, \sigma^2) p(\mathbf{f}|\mathbf{X}, \theta) d\mathbf{f}$$

- Log marginal likelihood

$$\log p(\mathbf{y}|\mathbf{X}) \propto -\frac{1}{2}\mathbf{y}^T (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{y} - \frac{1}{2} \log |\mathbf{K} + \sigma^2 \mathbf{I}| - \frac{N}{2} \log(2\pi)$$

- The parameters θ and σ^2 are estimated by using a gradient- based algorithm
  - Maximum Likelihood
  - Integration via Hybrid Monte Carlo

Gaussian Process Regression

- As we now have all components for determining the posterior distribution we can evaluate new predictive distributions of any new test data, conditioned on training results

$$p(f_*|\mathbf{x}_*, \mathbf{y}, \mathbf{X}, \theta, \sigma^2)$$

- Mean and variance are then given by

$$m(\mathbf{x}_*) = \phi(\mathbf{x}_*)^T \mu = \mathbf{k}_*^T (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{y}$$

$$v^2(\mathbf{x}_*) = \phi(\mathbf{x}_*)^T \Sigma \phi(\mathbf{x}_*) = k_{**} - \mathbf{k}_*^T (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{k}_*$$

Predictor Selection

- Two sources can contribute to streamflow predictability
  - Influence of initial catchment conditions: (antecedent streamflow, precipitation, temperature)
  - Effect of climate during the forecasting period: (climate indices)
- Here: Focus on catchment conditions
  - Predictor Group 1: $Q_{t-1}, Q_{t-2}, P_{t-1}, T_{\max t-1}, T_{\max t-2},$ and $T_{\min t-1}$
  - Predictor Group 2: $Q_{t-1}, Q_{t-2} P_{t-1}, T_{\max,t-1} T_{\max,t-2} T_{\min,t-1} \bar{P}_t \bar{T}_{\max,t} \bar{T}_{\min,t}$

    including long-term monthly averages

Performance metrics

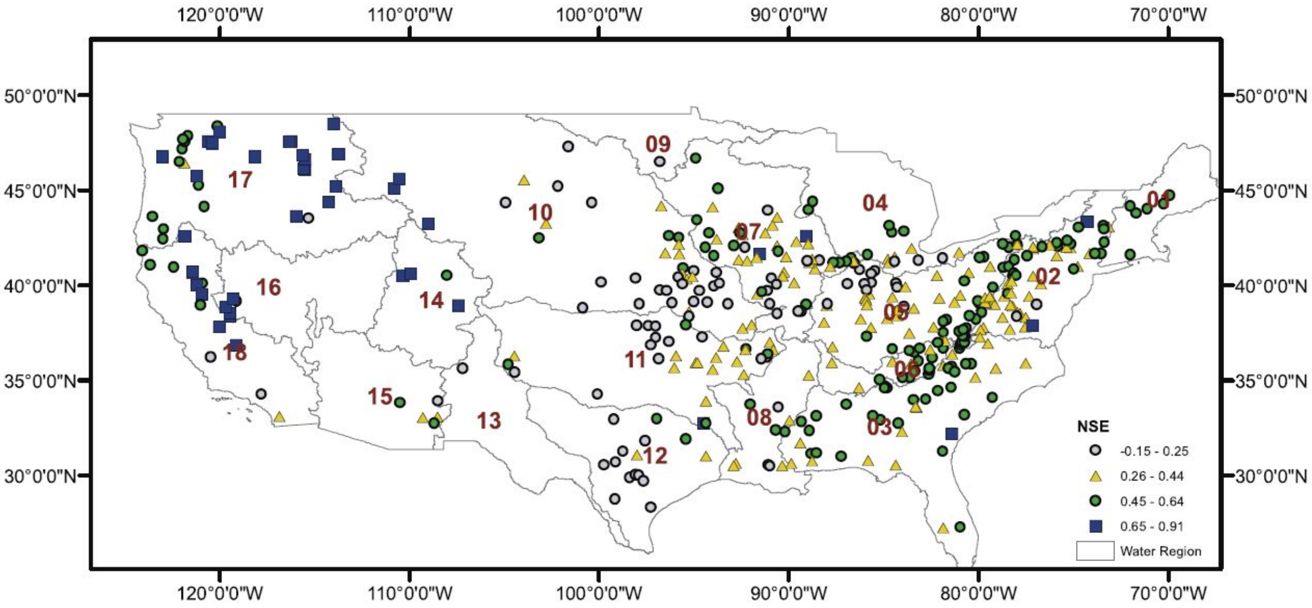
- Standard Nash-Sutcliffe efficiency (NSE)
  - Quantifies the skill of a model to explain streamflow variance
  - Sensitive to extreme values

$$NSE = 1 - \frac{\sum_{i=1}^{n}(Q_i - Q_{o,i})^2}{\sum_{i=1}^{n}(Q_{o,i} - \bar{Q}_o)^2}$$

- Mean cumulative error/ water balance error
  - ability of a model to correctly reproduce streamflow volumes

$$WB = 1 - \left| 1 - \sum_{i=1}^{n} Q_i \Big/ \sum_{i=1}^{n} Q_{o,i} \right|$$

Map of MOPEX stations



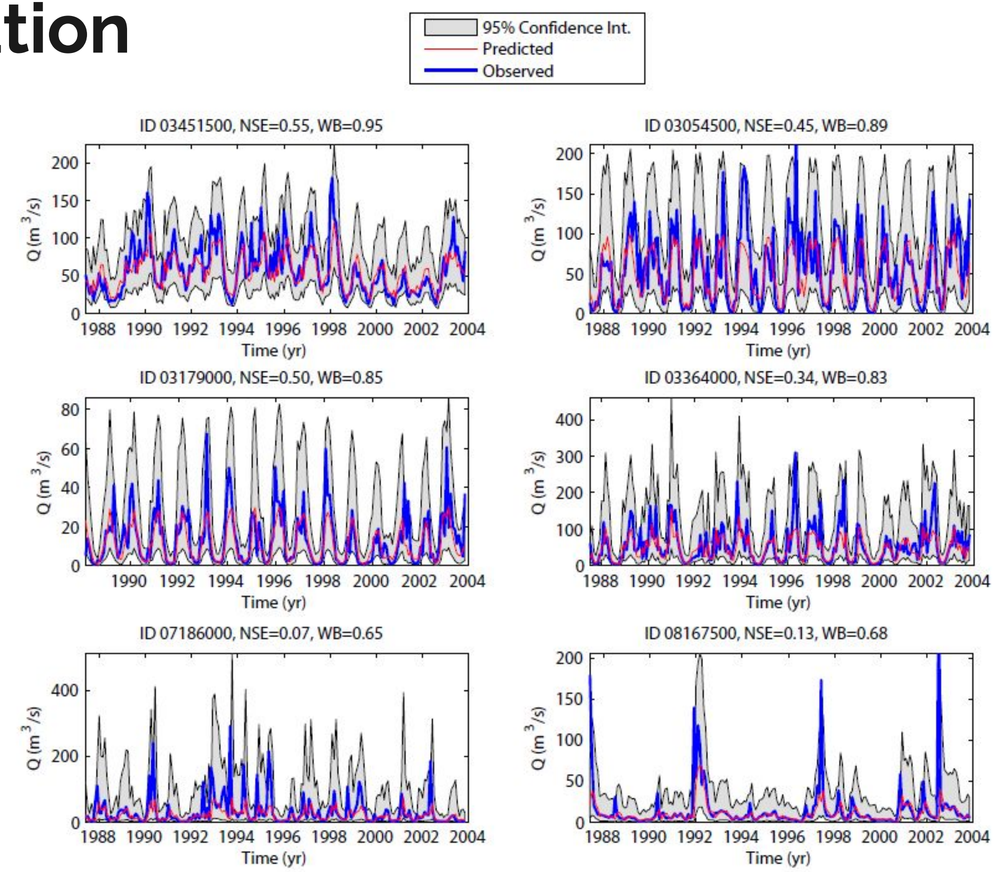
Performance validation

- 6 out of 12 MOPEX basins were selected to check performance
  - Basins range from very wet to very arid
- Annual Q/P (runoff-ratio)
- Mean annual potential evaporation/precipitation = PET/P (arridity index, evaporation index = 1-runoff ratio)

Station ID	Lon. (deg)	Lat. (deg)	Area (km^2)	Annual Q/P	Annual PET/P	Gauge Info
03451500	–82.58	35.61	2445	0.5	0.54	French Broad River at Asheville, NC
03054500	–80.04	39.15	2361	0.56	0.54	Tygart Valley River at Philippi, WV
03179000	–81.01	37.54	1024	0.42	0.76	Bluestone River near Pipestem, WV
03364000	–85.93	39.20	4419	0.37	0.83	East Fork White River at Columbus, IN
07186000	–94.57	37.25	2999	0.26	1.01	Spring River near Waco, MO
08167500	–98.38	29.86	3457	0.13	1.98	Guadalupe River near Spring Branch, TX

Performance validation

- NSE tends to improve, moving from dry to humid regions (higher runoff ratios and lower aridity index)
- GPR captures streamflow adequately, except for flashy flooding events



Performance validation - Model comparison

- Most stations show better results under GPR
- Four GPR underperformers belong to erratic regimes that are less predictable in general

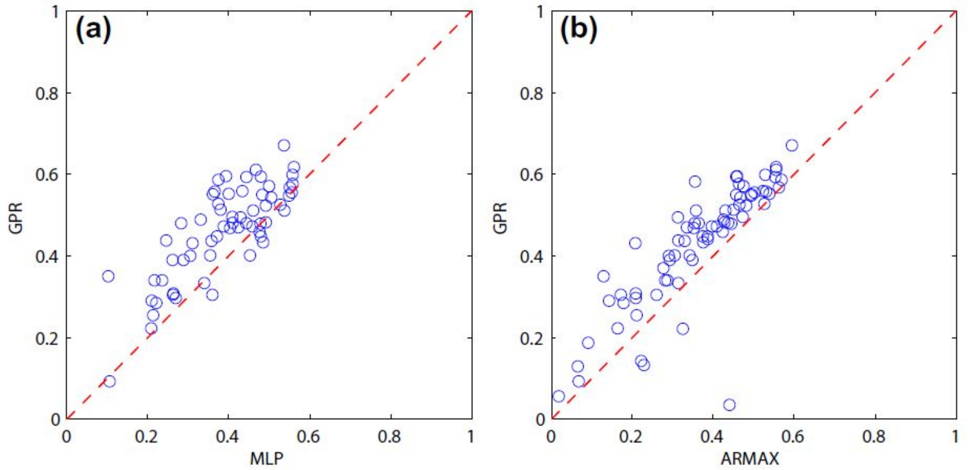


Fig. 3. Comparison of NSE obtained by (a) GPR and MLP and (b) GPR and ARMAX, for original MOPEX testing data.

Performance validation - extended/original MOPEX

- Ideally relatively stable and unchanged performance for testing and extended period
    - The lower the NSE, the more unstable the predictions.
    - Higher NSE seem to persist into extended period
- Possible explanations:
    - Erratic flow regime and general less predictable for low NSEs
    - Dry areas with sporadic rainfall
    - Anthropogenic impacts that alter watershed
    - Effect of nonstationarity (data-driven models trained on Historic data are no longer valid)

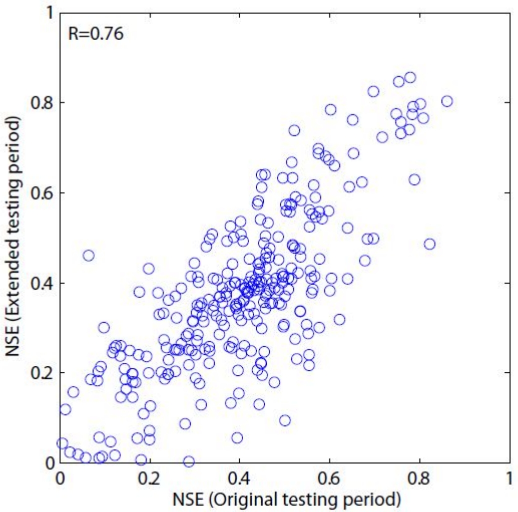
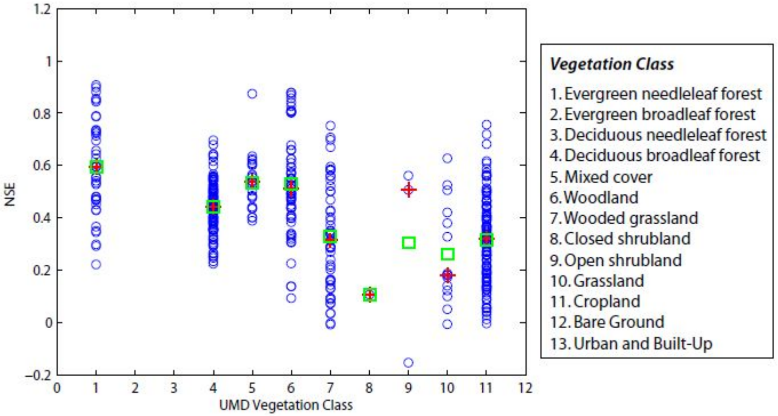


Fig. 4. Comparison of GP model performance on the original MOPEX testing data and those in the extended period (2004–2012).

Factors affecting GPR predictability

- Basins with best predictability tend to be energy-limited
- Basins with worst predictability tend to be water supply-limited regions in (semi-)arid regions



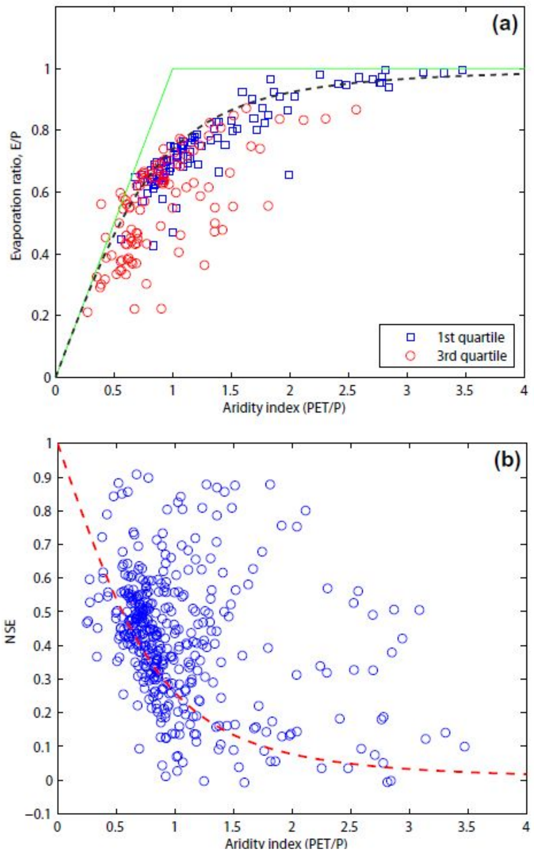


Fig. 5. (a) Budyko diagram for illustrating NSE similarity, where square and circle symbols correspond to NSE's <1st and >3rd quartiles, respectively. The horizontal axis is aridity index and vertical axis is evaporation ratio. Budyko curve is the gray dash line and the limit lines are in green; (b) NSE vs. aridity index.

Summary

- GP models for more than 400 MOPEX basins trained to perform one-month-ahead streamflow forecast.
- GPR mostly outperforms ARMAX and MLP
- Little sensitivity to different kernels
- Basins with best predictability tends to be energy-limited/ worst water supply-limited
- Bains in the Pacific Northwest and eastern U.S. generally higher predictability than basins located in the Midwest.