Capacity Building in Seasonal Hydrological Forecasting

Modeling and Machine Learning

AGRHYMET, Climate Regional Center for West-Africa and Sahel

@Arsène KIEMA

2025-10-06

Pedagogical Objectives

Learning outcomes

By the end of this module, participants will be able to:

- Understand the theoretical foundations of four key models: PCR, Ridge, Lasso, and Random Forest.
- Explain how these models handle collinearity, regularization, and overfitting.
- Train and evaluate each model in R using hydrological data.
- Interpret model coefficients, variable importance, and performance metrics.

Statistical and Machine Learning Models

Hydrological modeling often relies on relationships between predictors (e.g., rainfall, temperature, evapotranspiration) and target variables (e.g., streamflow, runoff).

Statistical and ML models can:

- Capture linear and nonlinear relationships.
- Handle multiple correlated predictors.
- Provide predictive tools for bias correction or forecasting.

Models covered today

Model	Type	Key idea	Regularization
PCR	Linear	Use principal components of predictors	Implicit
Ridge	Linear	Penalizes large coefficients (L2)	L2 penalty
Lasso	Linear	Performs variable selection (L1)	L1 penalty
Random Forest	Nonlinear ensemble	Combines multiple decision trees	Implicit

Reminder: Linear Regression

Linear regression assumes a linear relationship between predictors and the target:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon$$

Where:

- (Y) = target variable (e.g., streamflow)
- (X_i) = predictors (e.g., rainfall, PET, temperature)
- (_i) = coefficients
- ( ) = random error term

Reminder: Linear Regression

! Limitation

When predictors are highly correlated (collinearity), standard linear regression becomes unstable.

Principal Component Regression (PCR)

PCR combines:

1. **Principal Component Analysis (PCA)** $\rightarrow$ reduces correlated predictors to a few uncorrelated components.
2. **Linear Regression** $\rightarrow$ fits the target variable using these principal components.

Mathematically:

$$Y = \alpha_0 + \alpha_1 PC_1 + \alpha_2 PC_2 + \ldots + \alpha_k PC_k + \varepsilon$$

- (PC_k) are the first *k* principal components explaining most of the variance.

Principal Component Regression (PCR)

Advantages

- Solves multicollinearity.
- Reduces noise and dimensionality.
- Useful when predictors » observations.

Limitations

Principal components are linear combinations of predictors, hence less interpretable.

Ridge Regression (L2 Regularization)

Ridge regression penalizes large coefficients by adding an **L2** term to the cost function:

$$\text{Minimize } \sum (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

- ( ): regularization strength.
- Larger ( ) $\rightarrow$ stronger penalty $\rightarrow$ smaller coefficients.

Ridge Regression (L2 Regularization)

Advantages

- Stabilizes model coefficients under collinearity.
- Reduces overfitting while keeping all predictors.

Limitations

- Does not perform variable selection; all predictors remain in the model.

Lasso Regression (L1 Regularization)

Lasso adds an **L1** penalty to shrink some coefficients exactly to zero:

$$\text{Minimize } \sum (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- Performs both **regularization** and **feature selection**.

Lasso Regression (L1 Regularization)

Advantages

- Produces simpler, interpretable models.
- Selects the most influential variables.

Limitations

When predictors are highly correlated, Lasso tends to select one and discard the others.

Random Forest (RF)

Random Forest is a **nonlinear ensemble model** built from many decision trees.

Each tree: - Uses a **bootstrap sample** of the data.
- Splits variables randomly at each node.

The final prediction is the **average** (for regression) of all trees:

$$\hat{Y}_{RF} = \frac{1}{N_{trees}} \sum_{t=1}^{N_{trees}} \hat{Y}_t$$

Random Forest (RF)

Advantages

- Captures complex nonlinear relationships.
- Robust to noise and correlated predictors.
- Provides variable importance measures.

Limitations

- Less interpretable than linear models.
- Requires more computational time and tuning.

Summary Table

Model	Type	Handles Collinearity	Variable Selection	Nonlinear	Regularization
PCR	Linear	Yes	No	No	Implicit
Ridge	Linear	Yes	No	No	L2
Lasso	Linear	Yes	Yes	No	L1
Random Forest	Nonlinear	Yes	Yes (implicit)	Yes	Implicit

THANK YOU FOR YOUR ATTENTATION