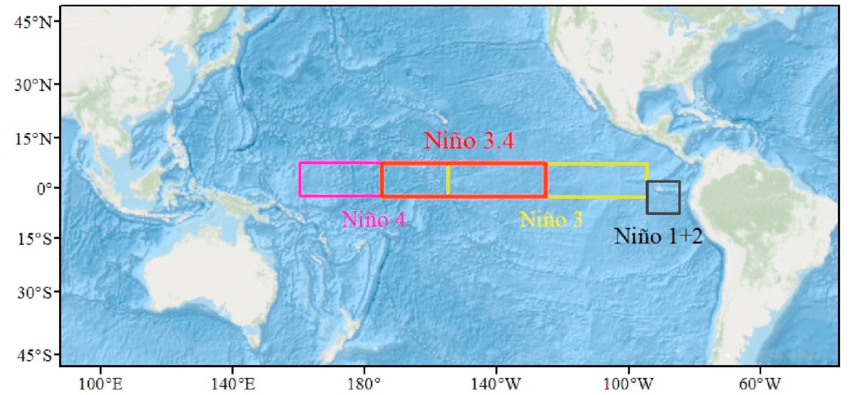
Machine learning metasystem for robust probabilistic nonlinear regression-based forecasting of seasonal water available in the US west

* <https://www.nrcs.usda.gov/sites/default/files/2022-11/A%20Machine%20Learning%20Metasystem%20for%20Robust.pdf>
* 3 considerations
  + Institutional requirements around what is and what is not a logistically feasible applied science and engineering solution to prediction task
  + Largely successful decades long track recording of existing PCR (principal components regression) based forecast system
  + PCR system’s consistency with the large body of environmental and geophysical science knowledge around water resources and their prediction point in combination to a solution that involves building upon the existing PCR framework
    - Aka multicollinearity mitigation and dimensionality reduction through independent signal extraction from specific known classes of geophysical predictor datasets, followed by phenomenological modeling process relating selected signals to the predictand through some form of regression like input output mapping, along with automated process for optimally choosing which candidate input variables, and signal sderived from them in the initial data preprocessing step, to retain the final model
* Water supply forecasting (WSF) started 1920s
  + Manual snow measurements
* Modern methods use process based or data driven approaches
  + Process simulation approaches = mathematical models that deterministically represent the large number of geophysical and biophysical processes (forest and crop evapotranspiration, snowpack accumulation and melt, rainfall and snowmelt infiltration, groundwater interactions, etc) and corresponding environmental parameters that control river runoff production for a given watershed
  + Data-driven models = do not explicitly represent underlying physical processes, instead using empirical fits between inputs like snowpack, precipitation, and climate data, and outputs like seasonal river runoff volume
    - Consist of statistical models, typically multiple linear regression or principal components regression using heuristic prediction bounds based on out-of-sample (cross-validated) standard error as a measure of predictive error variance
    - Regression predictand = April-july aggregated river flow or reservoir inflow volume
    - Forecasts begin to be issued in January or February and may continue into may or June
    - Predictor variates typically consist of mountain snowpack measurements taken just before the forecast is made
    - Also: M-regression, partial least squares regression
* Process simulation models provide physical insights and diagnostics
* Data driven models cheaper and can be computationally more reliable and faster, more amenable to incorporating new predicter data types (newly discovered climate indices, match or exceed the prediction performance or process simulation models, provide more reasonable prediction uncertainty info)
* Neural networks and support vector machines = ML used for hydrological applications
* Quantitative prediction uncertainty estimates = core product of all modern water, weather, and climate prediction systems, but aren’t integrated in many ML methods
* WSF predictor variates
  + Snowpack
  + Soil moisture
  + Precipitation data
  + …within or near watershed area (typically remote high elevation environmental measurement stations with automated data collection and telemetry
  + Antecedent streamflow observations
  + Indices of interannual climate variation (eg ENSO)
  + Output from process based snow models
  + Snow estimates from airborne or satellite remote sensing
  + Selection of specific input values based on hydrologist expertise, includes factors like spatial proximity, incorporating redundancy through multiple partially correlated SNOTEL sites in the event of sensor or telemetry failures during forecast operations, capturing local scale environmental heterogeneity across watershed using variety of SNOTEL sites, etc
* For each predictand, each forecast issue date has its own regression model and corresponding set of predictor variates, which evolve over the forecast season
* Data driven WSF system requires
  + Methods for addressing predictor multicollinearity
  + Identifying multiple input signals and potential WSF predictive value
  + Objective means for identifying the most promising predictor variables from a pool of broadly reasonable candidates
  + Relating these to forthcoming water supply availability using regression-like model
* Dimensionality and multicollinearity addressed using principal component analysis (PCA) data preprocessing
  + PCA = pattern recognition technique that compresses the info content of a large dataset into a series of mutually uncorrelated modes that efficiently concentrate the total dataset variance
* Model selection problem addressed with multi method ensembles
  + More diverse the models, the better

NINO SST

* El nino = happens in equatorial pacific ocean, 5 consecutive 3 month running mean of SST anomalies in the nino 3.4 region that is above/below the threshold of 0.5 degrees C
* For development and persistence of deep convection (enhanced cloudiness and precipitation) in the tropics, local SST must be 28 degrees C or greater
  + Once pattern of deep convection altered due to anomalous SSTs, the tropical and subtropical atmospheric circulation adjusts to new pattern of tropical heating, resulting in anomalous patterns of precipitation and temperature that extend beyond the region
  + SST anomaly of 0.5 C in Nino 3.4 is sufficient to reach this threshold from late march to mid June
  + A larger anomaly (up to 1.5 C) in November-January required to reach the threshold to support persistent deep convection in that region
* SST anomaly of -0.5 C in Nino 4 sufficient to bring water temperatures below the 28 C threshold, which would result in a significant westward shift in the pattern of deep convection



Source: <https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.researchgate.net%2Ffigure%2FOutline-of-the-four-Nino-regions-over-the-tropical-Pacific-Ocean-and-the-Nino-34-region_fig2_348627118&psig=AOvVaw0sdZQaHOSdSOhEi-Nf26SD&ust=1699185966324000&source=images&cd=vfe&opi=89978449&ved=0CBEQjhxqFwoTCMia4pGnqoIDFQAAAAAdAAAAABAD>

A table with numbers and lines

Description automatically generated

CSV file after converting from txt

A table with numbers and letters

Description automatically generated

Created new index in datetime

A graph showing the number of numbers and the number of the number of numbers

Description automatically generated with medium confidence

SST in C for each Nino region throughout the years

A graph with numbers and lines

Description automatically generated

Anomalies in C for each Nino region throughout the years

### ONI Oceanic Niño Index

* measure is used as an indicator of the [El Niño–Southern Oscillation](https://www.ncei.noaa.gov/access/monitoring/enso/technical-discussion) phenomenon. Warm (El Niño) and cold (La Niña) phases are defined as a minimum of five consecutive ONI values surpassing a threshold of +/- 0.5°C.
* A screenshot of a whiteboard

  Description automatically generated

### Southern Oscillation Index

* Standardized sea level pressure differences between Tahiti and Darwin, Australia. This measure is used as an indicator of the [El Niño–Southern Oscillation](https://www.ncei.noaa.gov/access/monitoring/enso/technical-discussion) phenomenon. The index is negative when there is below-normal air pressure at Tahiti and above-normal air pressure at Darwin, and vice versa when the index is positive. Periods of negative values coincide with El Niño and positive values coincide with La Niña.
* A screen shot of a graph

  Description automatically generated

### Madden-Julian Oscillation (MJO) Pentad Indices

* The Madden-Julian Oscillation is an eastward moving weather pattern with a typical period of 30 to 60 days. The pentad indices are normalized projections of pentad velocity potential on patterns from extended empirical orthogonal function analysis on historical reference data from 1979 to 2000.

### Pacific North American (PNA) Index

* large-scale weather in the atmospheric circulation over the Pacific Ocean and North America. The index is the projection of the air pressure field on a particular mode from empirical orthogonal function analysis of reference data from 1950 to 2000.
* Influences regional weather by affecting the strength and location of the East Asian jet stream and the weather it brings to North America.
* A colorful lines on a white background

  Description automatically generated

### Pacific Decadal Oscillation (PDO) Index

* a climate pattern of the Pacific Ocean that is characterized by warm and cool phases in sea surface temperature. It is similar to the El Niño–Southern Oscillation but has a longer time scale, with phases that can persist for 20 to 30 years. The index is calculated from projecting sea surface temperatures on the first principal component of reference data from 1900 to 1993.
* A screen shot of a graph

  Description automatically generated