# The Statistical Evaluation of the ESPAM2 Model

MAXINE DAKINS

ENVIRONMENTAL SCIENCE PROGRAM AND
IDAHO WATER RESOURCES RESEARCH
INSTITUTE

UNIVERSITY OF IDAHO

# The Statistical Evaluation of the ESPAM2 Model

Purpose:

- Recommend methods to statistically evaluate the fit of the ESPAM2.
- Identify appropriate statistical measures of model fit that will allow the two year evaluation period (2009-2010) to be compared to the 29 year calibration period (1980-2008).
- In order to evaluate the fit of the model to data that were not used in model calibration.

# Root Mean Squared Error (RMSE)

A common statistical measure of model fit is the root mean squared error (RMSE) (Hill and Tiedeman).

The RMSE is calculated from the objective function for the calibration, the sum of the squared errors (SSE).

RMSE = Root mean squared error = sqrt(SSE/df)

- where
  - df = n – p
  - n = number of data points and other points fixed points used in the calibration
  - p = number of parameters fitted in the regression

# Root Mean Squared Error

- The advantage of the RMSE over the SSE is that by taking the square root, the value has the same units as the data.
- Smaller RMSEs result from smaller SSEs and indicate a better model fit.
- Larger RMSEs indicate a worse model fit.

# Root Mean Squared Error

- Could compare the RMSE for the evaluation period against the RMSE from the calibration period
- An issue is that the comparison of two points is difficult
- One of them will be higher than the other even if the actual underlying fit is equivalent.
- Validation RMSE > Calibration RMSE 50% of the time if fit is equivalent.

# Root Mean Squared Error

- In addition, the RMSE of the model fit may vary through time
- For example, if the model fit is related to periods of wet and dry weather or any other variable that changes through time.
- So, need to have measures of model fit through the calibration period that capture the variability of the fit through time

# Root Mean Squared Error

- Could use multiple two year periods throughout the calibration time period
  - 1980 and 1981
  - 1981 and 1982
  - 1982 and 1983
  - and so on.
- These values are not independent since most years will be used in the calculation of two RMSEs.
- This approach would result in 28 RMSE values for the calibration period.

# Root Mean Squared Error

- Could calculate the RMSE for non-overlapping, two year periods throughout the calibration period.
  - 1980 and 1981
  - 1982 and 1983
  - 1984 and 1985
  - and so on.
- These values will be independent of each other.
- This will result in 14 RMSE values for the calibration period.
- This is the first recommended approach

# Root Mean Squared Error

- Then the RMSE for the evaluation period (2009 – 2010) should be compared to the distribution of RMSE values from the calibration period.
- An evaluation RMSE that falls within the range of calibration RMSE values would be considered acceptable and not invalidating the model.
- An evaluation RMSE falling outside of the range of calibration RMSEs on the high side might be considered troubling.

# Root Mean Squared Error

- A positive characteristic of the RMSE is that it is consistent with the objective function, they are both based on the sum of squared errors.
- Downside is that outliers or extreme deviations between the model and the measurement have a large influence as they are squared before being summed.
- This means that a small number of large deviations can exert a strong influence on the outcome of the calculation.

# Median Absolute Deviation

- Second approach is the Median Absolute Deviation
- The median absolute deviation (MAD) is a robust measure of statistical dispersion (Helsel and Hirsch).
- Robust means that it is still a good measure even if assumptions are not met
- It is unaffected by a small number of outliers or extreme deviations between the model and the measurement.

# Median Absolute Deviation

- The median of a set of data values is the 50$^{th}$ percentile of the data, the value which exceeds 50% of the values and is exceeded by 50% of the values.
- Sort the list of values from smallest to largest.
- With an odd number of values, there will be a unique median.
- With an even number of values, it is the average of the two middle values.

# Median Absolute Deviation

- The median absolute deviation is the median of the absolute values of the deviations from the data's median.
- $MAD = median\ (|X_i - median(X_j)|)$
- Where
  - $X_i$ = the $i^{th}$ data point in the set
  - $median(X_j)$ = the median of all of the data values in the set
  - $||$ = the absolute value of the deviations

# Median Absolute Deviation

- Because it is the middle value of a sorted list of data, it is unaffected by a few large values at the high end of the data set.
- So, if some unusually large deviations between the model and the measurement exist, the median of the absolute values of the deviations will not be affected.

# Other Measures to Consider

- Other approaches could include:
- The Coefficient of Determination ($R^2$)
  - Represents the percent of the variability in the data explained by the model
  - Ranges from 0 to 100%
  - Also based on the SSE so similar information as the RMSE
- The Interquartile Range
  - The difference between the $75^{th}$ percentile and the $25^{th}$ percentile
  - Similar characteristics as the Median Absolute Deviation

# Recommendations

- Recommended approach:
  - Use both the RMSE and MAD as the measures of model fit
  - Calculate them for successive, non-overlapping two year periods throughout the calibration period.
  - Compare the value for the evaluation period to the set of values from the calibration period.
  - An evaluation value that falls within the range of calibration values would be considered acceptable and not invalidate the model.
  - An evaluation value falling outside of the range of calibration values on the high side would be considered troubling and might require additional thought.

# References


- Draper, N.R. and H. Smith. 1981. Applied Regression Analysis. Wiley and Sons, NY, NY.
- Helsel, D. R. and R. M. Hirsch. 1993. Statistical Methods in Water Resources. Elsevier, NY, NY.
- Hill, Mary C. and Claire R. Tiedeman. 2007. Effective Calibration of Environmental Models with Analysis of Data, Sensitivities, Predictions and Uncertainty.