

Use of Artificial Neural Networks for Modeling Indicator Organisms in a Drinking Water Supply Watershed

Diane M.L. Mas

David P. Ahlfeld

Department of Civil and Environmental Engineering

University of Massachusetts Amherst

December 5, 2003

Pathogens and Indicator Organisms

- Pathogens are a leading source of impairment in US and MA surface waters
- Multiple and Complex
 - Human
 - Animal
 - Natural Organic Substrates
- Indicator Organisms - surrogate parameter

Modeling Indicator Organisms Motivation

- Predictive tool
 - water supply protection
- Understanding process
 - insight into biological, physical, chemical processes
 - watershed management

Artificial Neural Networks

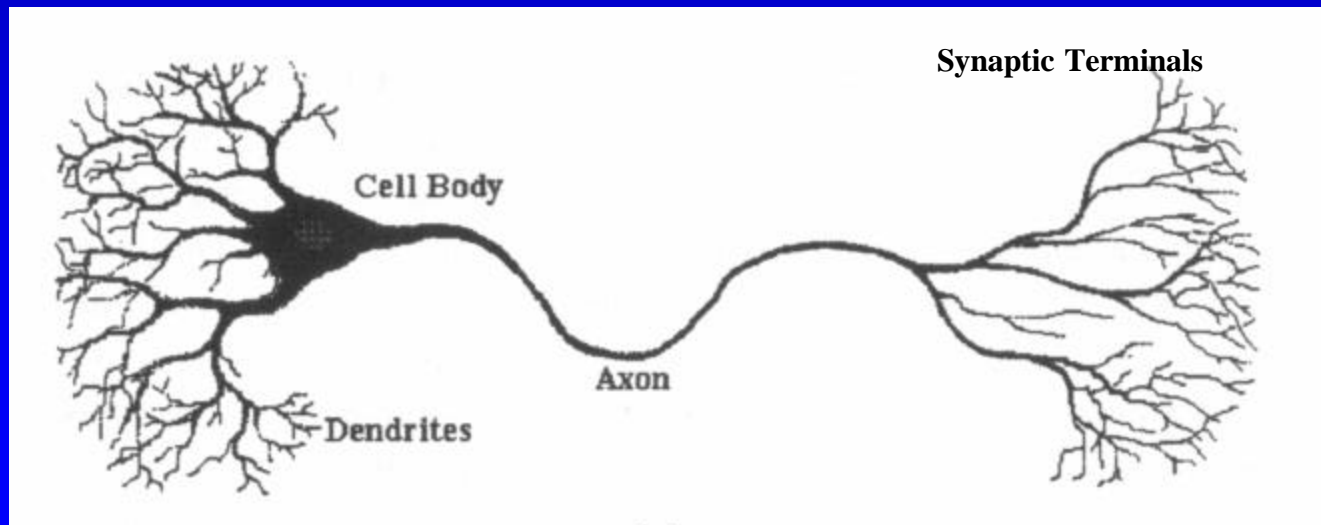
- Based on Structure of Biological Neural Networks
- Pattern recognition or function approximation
- Similar to statistical methods, but more flexibility in functional form

ANNs in Water Quality Modeling

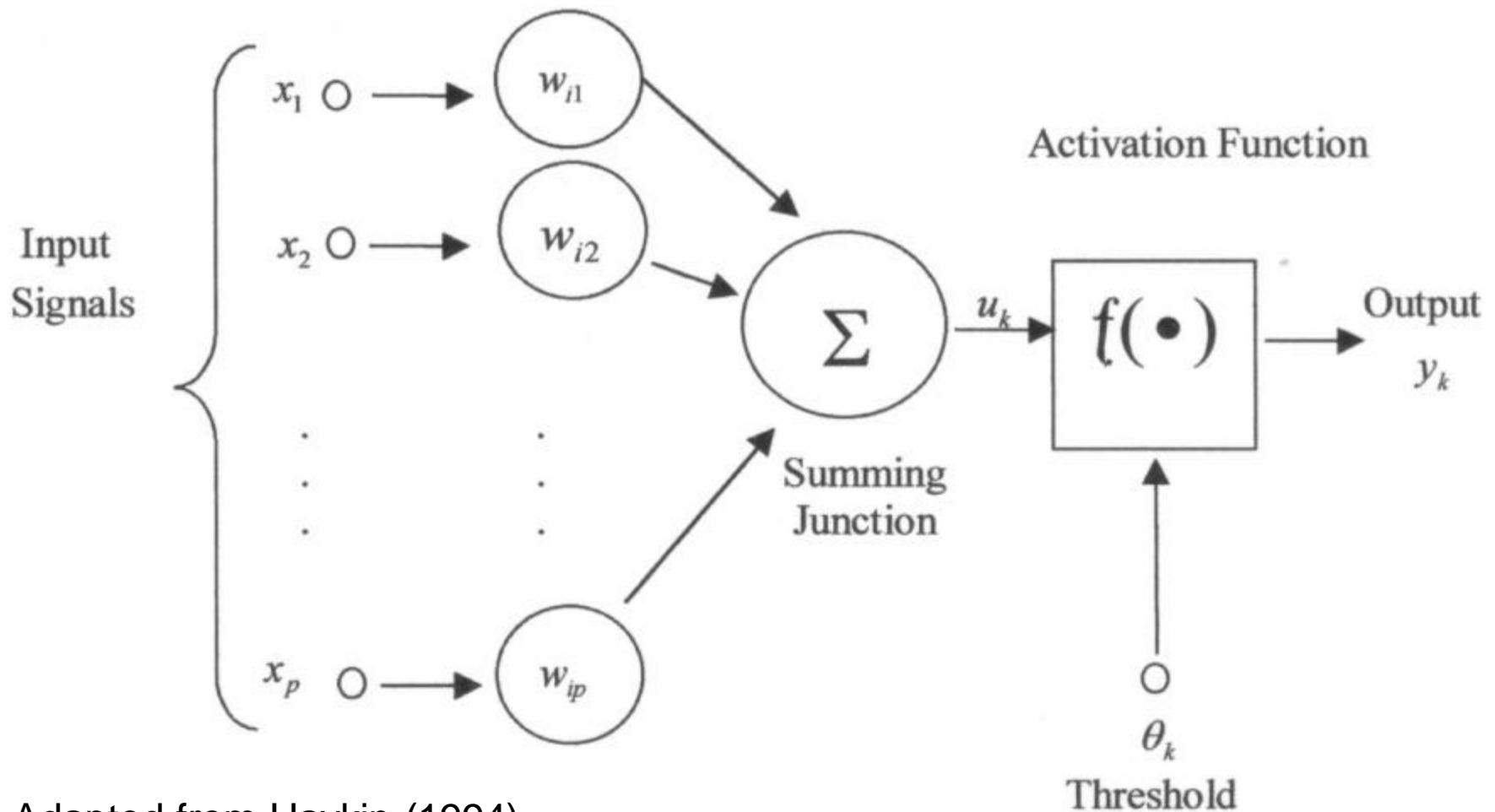
- Use in water resources in 1990s
- Attractive for complex, nonlinear problems
- Rainfall-runoff and flow modeling
- Limited biological/pathogen modeling

Biological Neuron

- **Dendrites** - Receive information from other neuron and transmit to cell body
- **Cell Body** - Collects and sums incoming info, >threshold --> action potential
- **Axion to Synaptic Terminal** --> action occurs

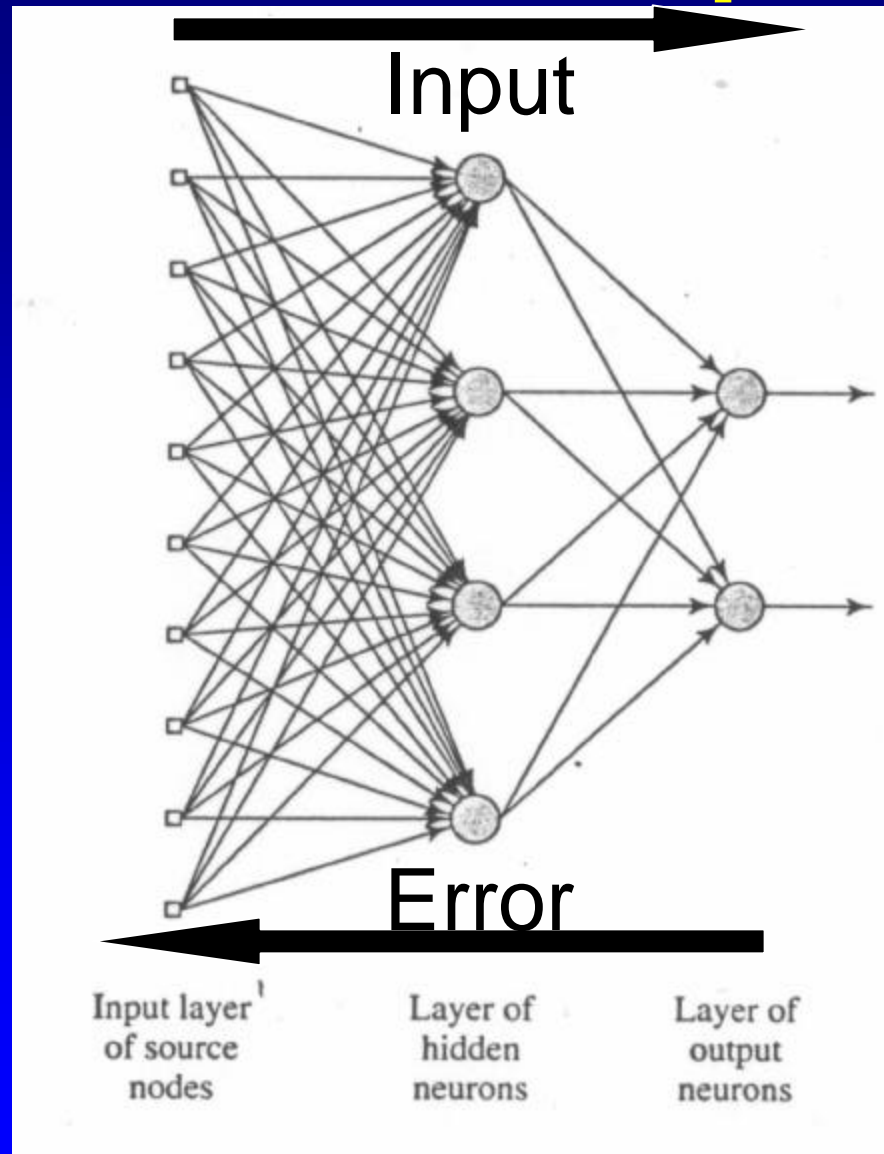


Schematic ANN Architecture



Adapted from Haykin (1994)

Feedforward - Backpropagation



Training Process

Backpropagation

- Training data presented - input and target
- Error between target output and simulated output calculated
- Error propagated back through network
- Connection weights updated to minimize error

Why Investigate ANN models?

- Increasing in use
- Widely available
- Relatively new; “black box” character
- Need to understand potential and limitations

Research Objectives

- Desire to evaluate ANNs as modeling tool for practitioners
- Provide practical guidance
- Three problematic areas
 - Input data preparation and model setup options
 - Applicability under different temporal or physical conditions
 - Strategies for utilization in unmonitored or under-monitored watersheds

Research Phases

- **Phase I**

- Effect of input data preparation and model architecture

- **Phase II**

- Transferability (temporal and spatial)

- **Phase III**

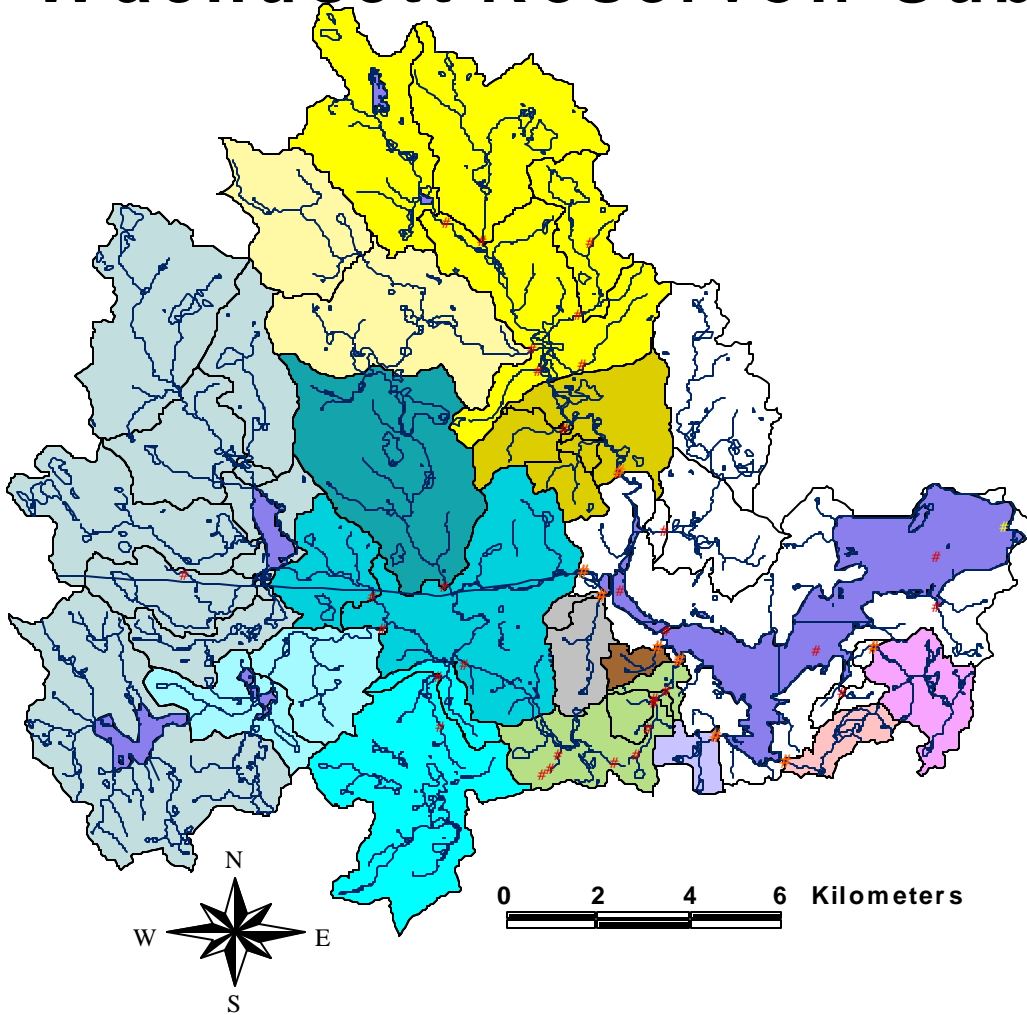
- Link with simple process-based models

Research Design and Methods

- Research Site
 - Wachusett Reservoir Watershed (303 km²)
 - MetroBoston water supply
 - Gates Brook Subwatershed (8.2 km²)



Wachusett Reservoir Subwatersheds



- # Stations with Streamflow Gaging
- # Sampling Stations
- Streams
- Subwatersheds
- Waterbodies
- French Brook
- Malagasco Brook
- Muddy Brook
- Gates Brook
- West Boylston
- Malden Brook
- Quinapoxet River
- Asnebumskit Brook
- Chaffins Brook
- Quinapoxet River
- Trout Brook
- Worcester Water Supply
- Stillwater River
- East Wachusett Brook
- Northern Stillwater
- Southern Stillwater

Available Data

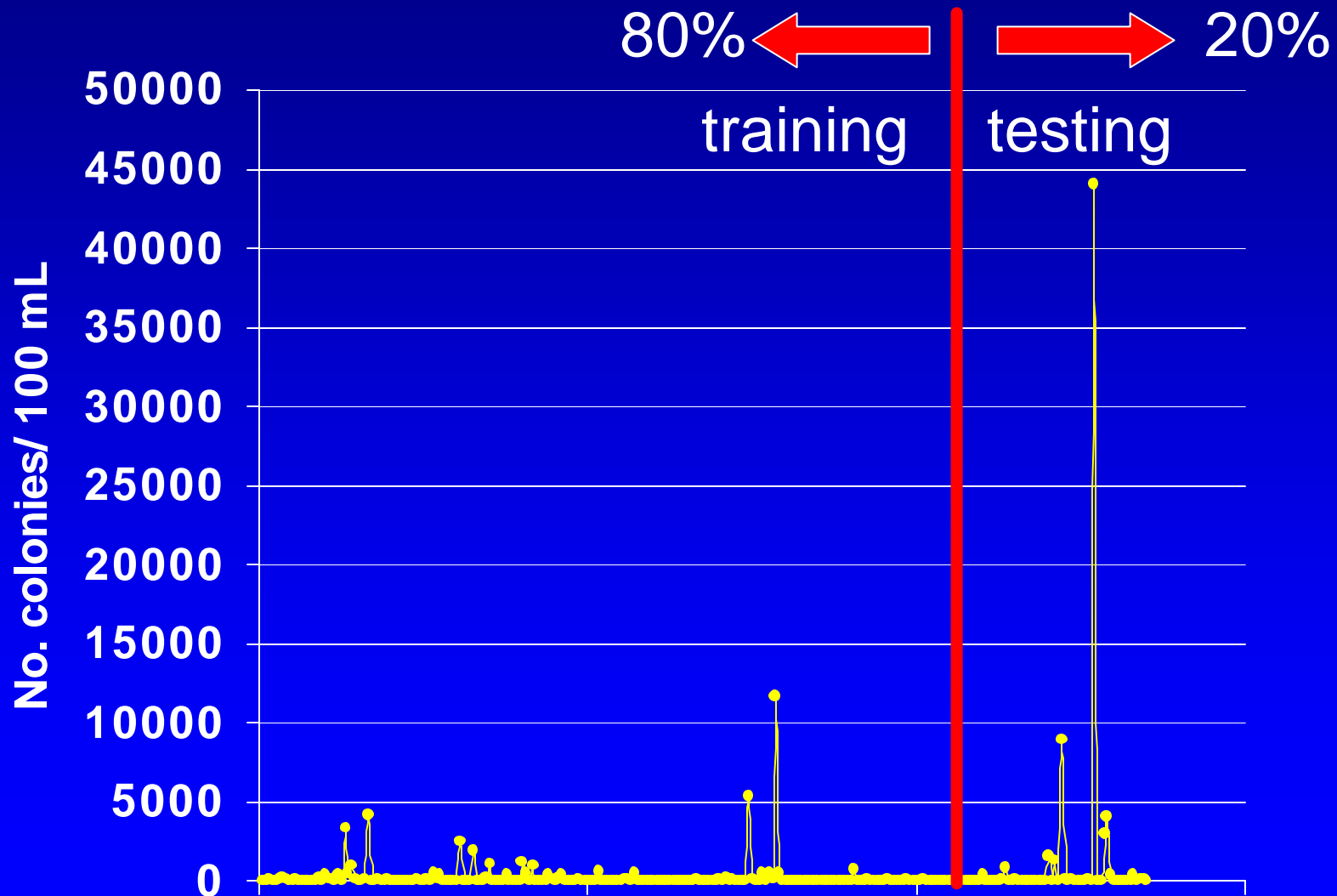
Gates Brook 1995 - 2000

- Fecal coliform
- Water Temperature
- Conductivity
- Instantaneous Streamflow
- Temperature and Precipitation Data (daily)

Preliminary Model Development

- Exploratory Data Analysis
 - Log normally distributed
 - Fecal Coliform
 - Streamflow
 - Conductivity
 - Precipitation
 - Significant ($P < 0.05$) but weak to moderate ($|r| < 0.5$) correlation between fecal coliform and
 - Streamflow
 - Conductivity
 - Precipitation

Distribution of Fecal Coliform Data 1995-2000



Effect of Input Data Selection and Preparation

- Random vs. Ordered Input Data
 - Capturing widest range of training values
- Data Normality
 - Better performance with normally distributed data when mean square error is used for error function?
- Sensitivity to Input
 - Better performance using only parameters that correlate to fecal coliform?

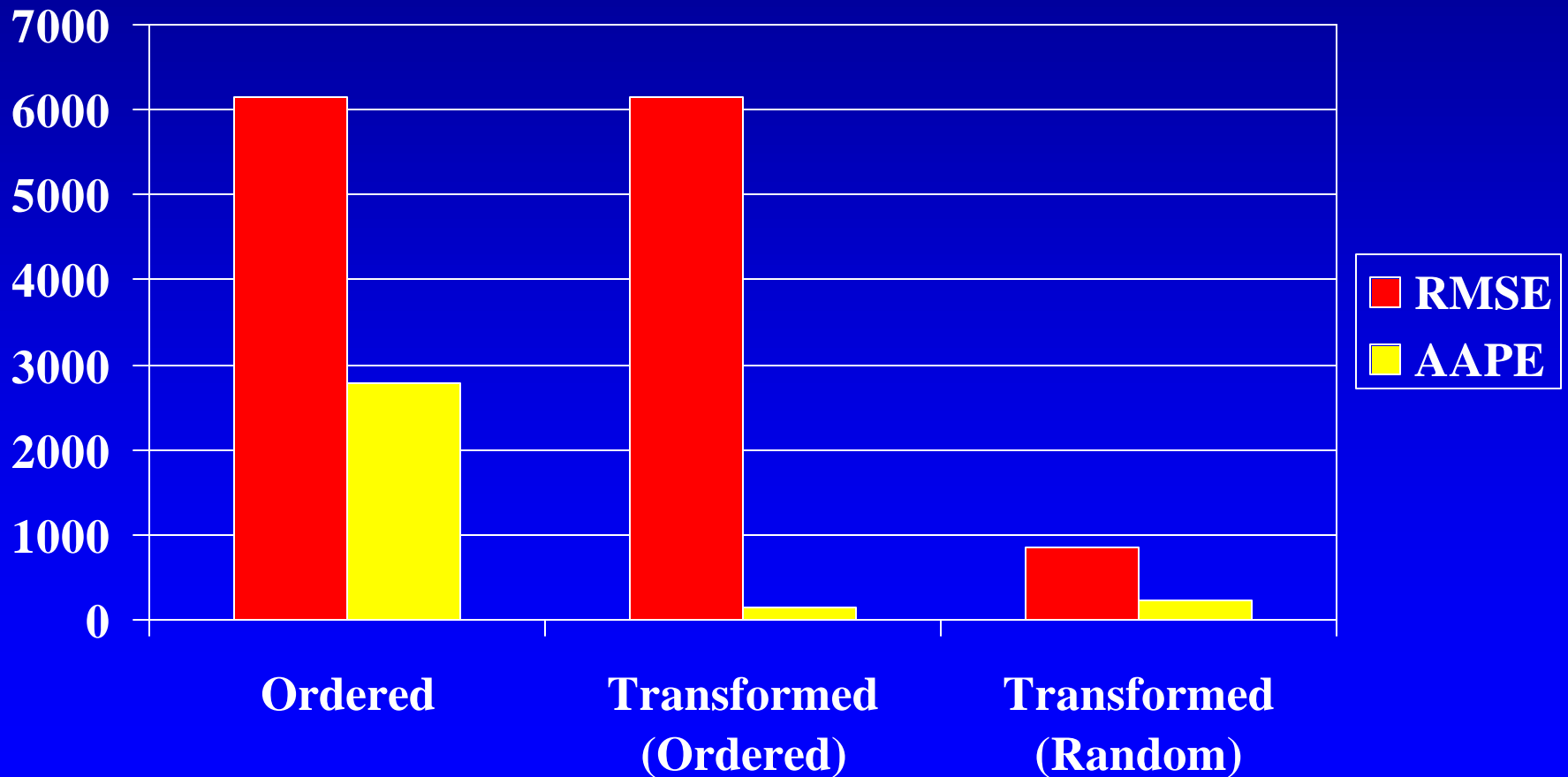
Development of ANN

- Commercial software = MATLAB Neural Network Toolbox
- Multilayer Feedforward ANN
- 3-Layer Structure = Input, Hidden Layer, Output (7:3:1)
- Optimization Method = Backpropagation
- Activation Function = Logistic Function
- Error Function = Mean Square Error

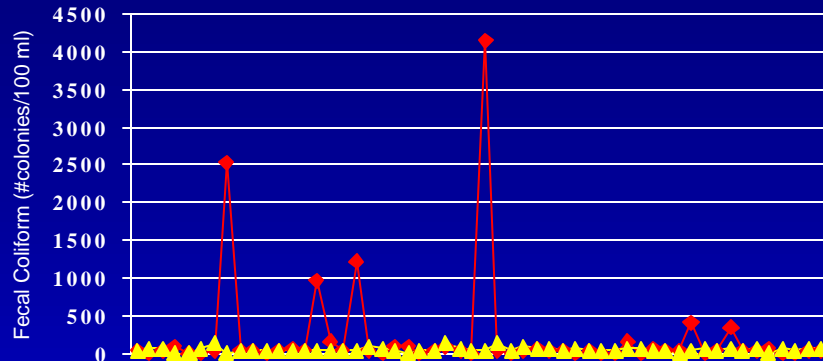
Model Performance Measures

- Root Mean Square Error (RMSE)
- Average Absolute Percent Error (AAPE)
- Visual Comparison

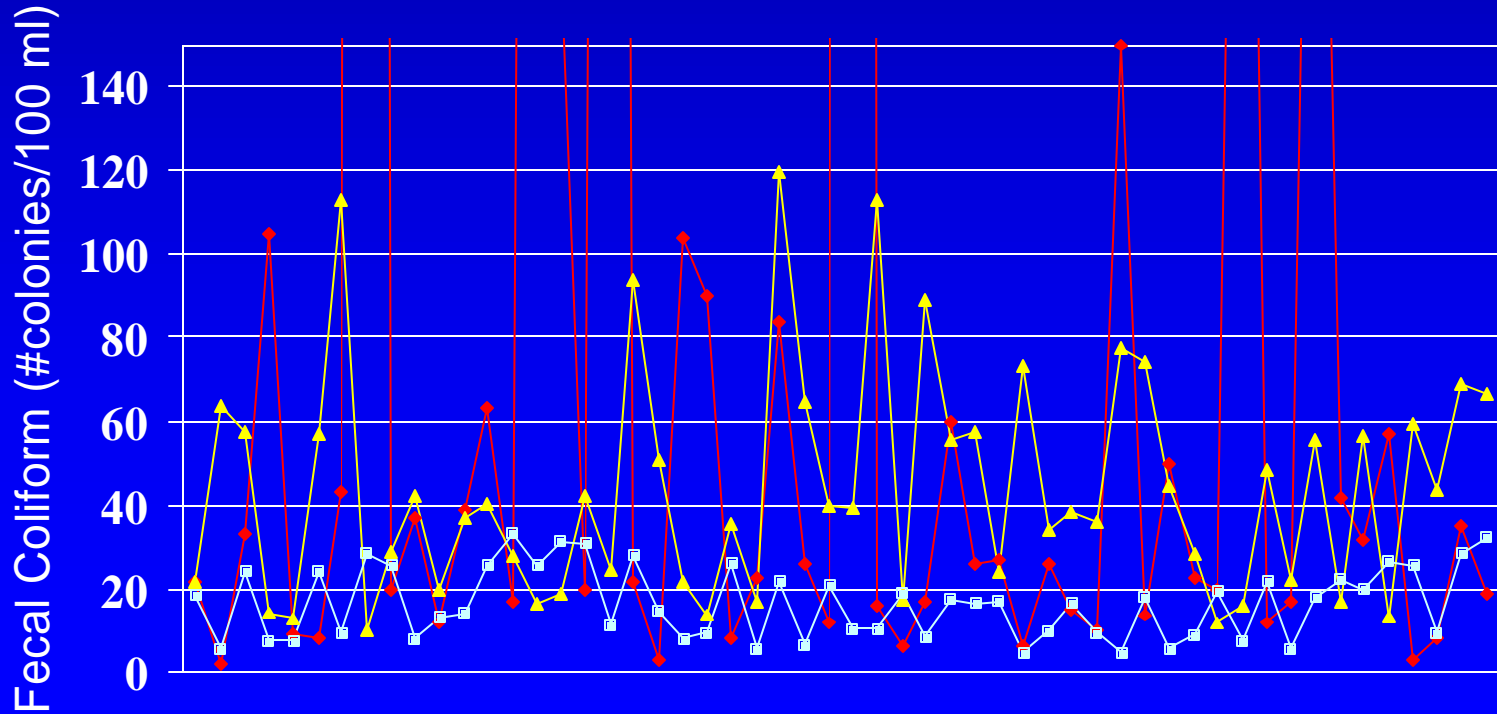
Model Performance Evaluation



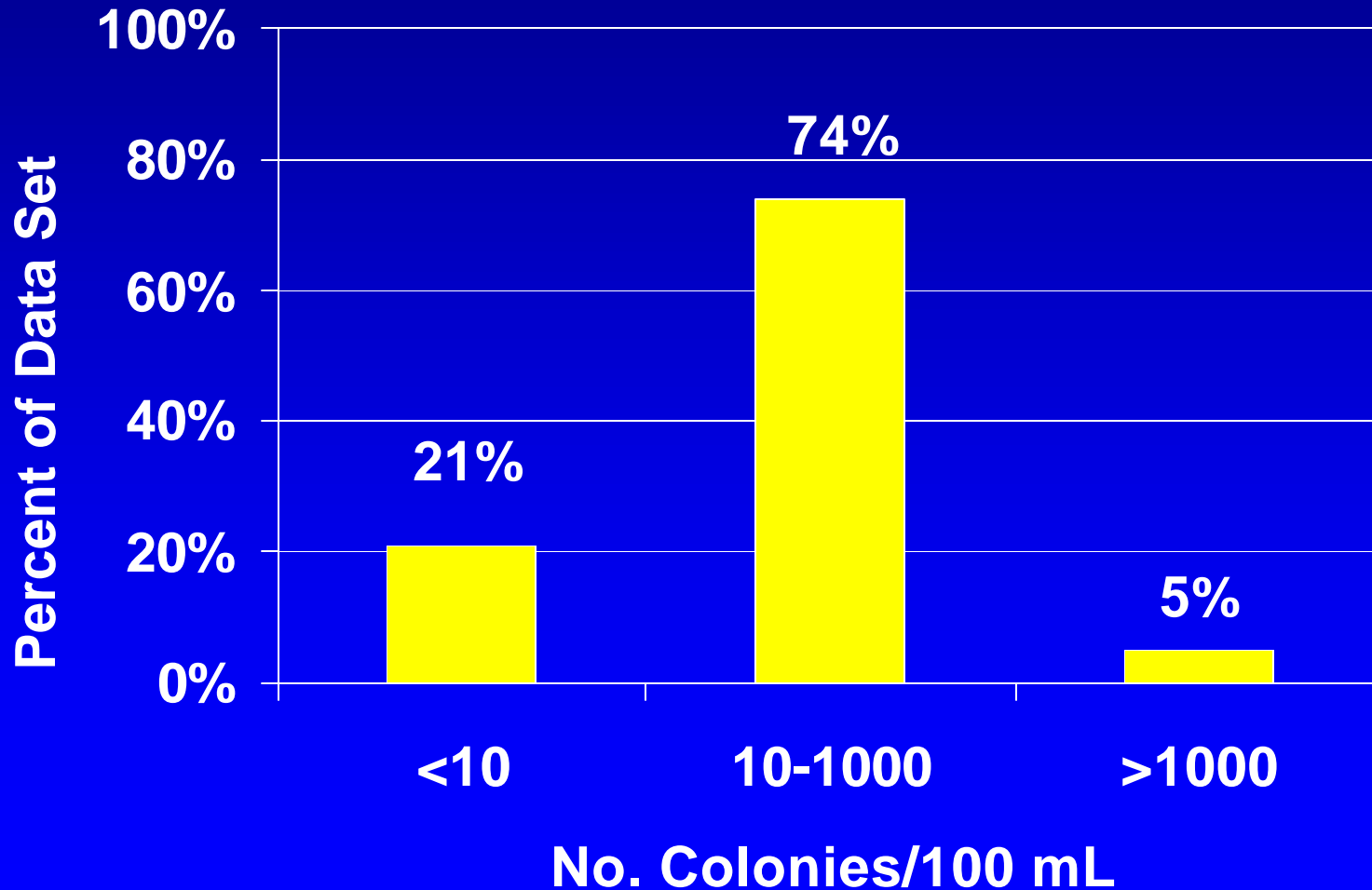
Visual Comparison



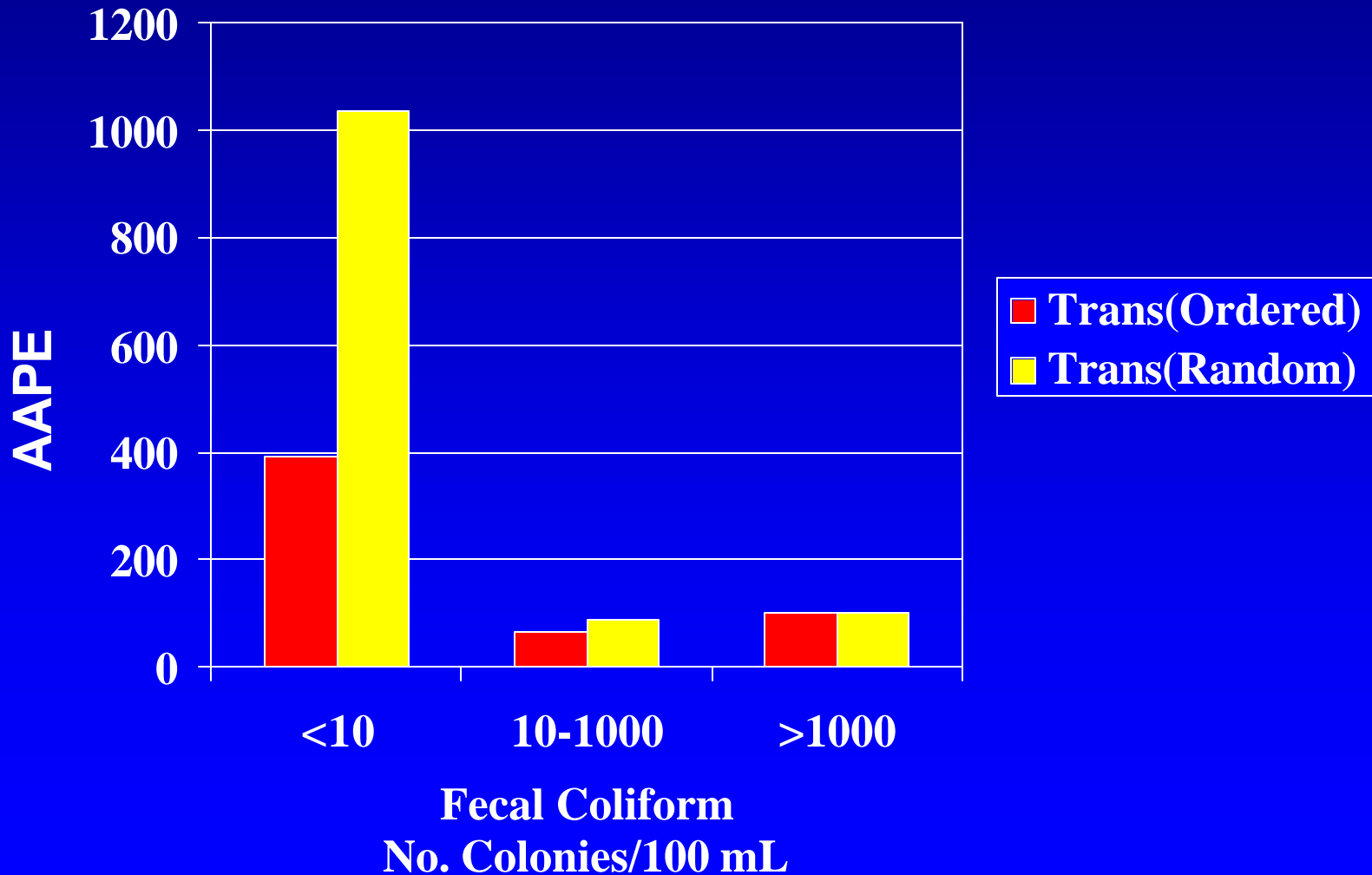
—◆— Target —▲— Trans(Random) —■— Trans(Ordered)



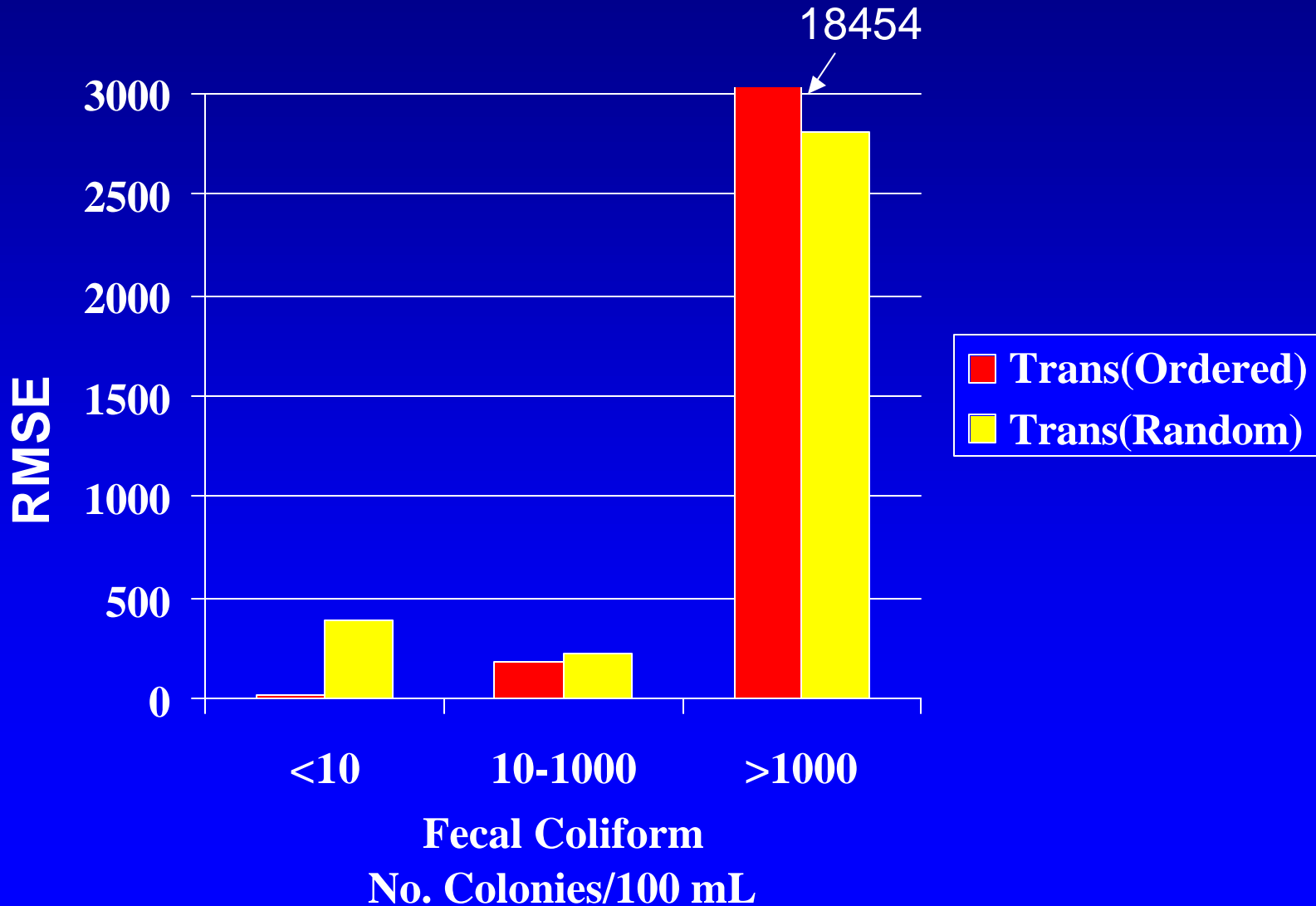
Distribution of Fecal Coliform Data 1995-2000



Model Performance Evaluation

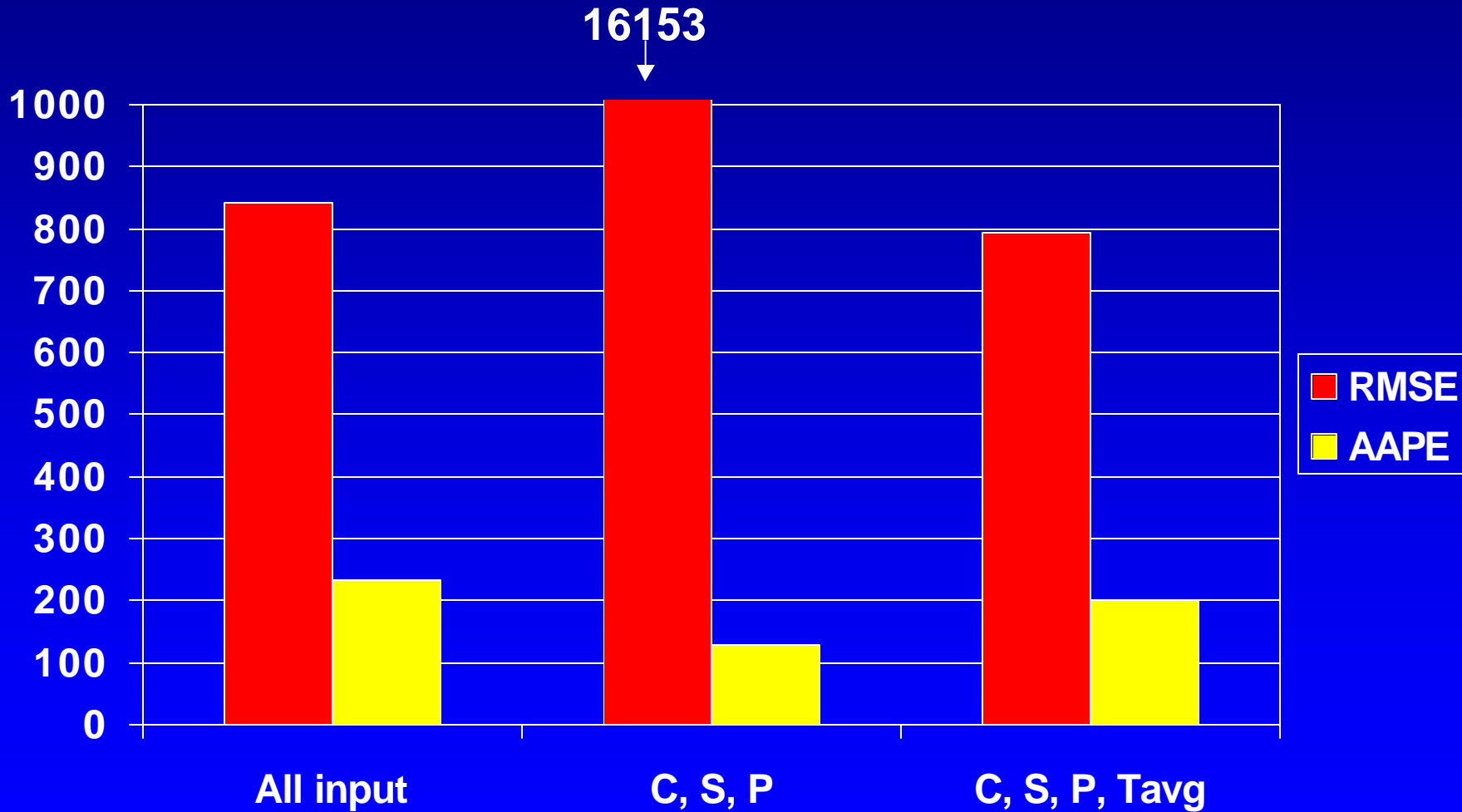


Model Performance Evaluation



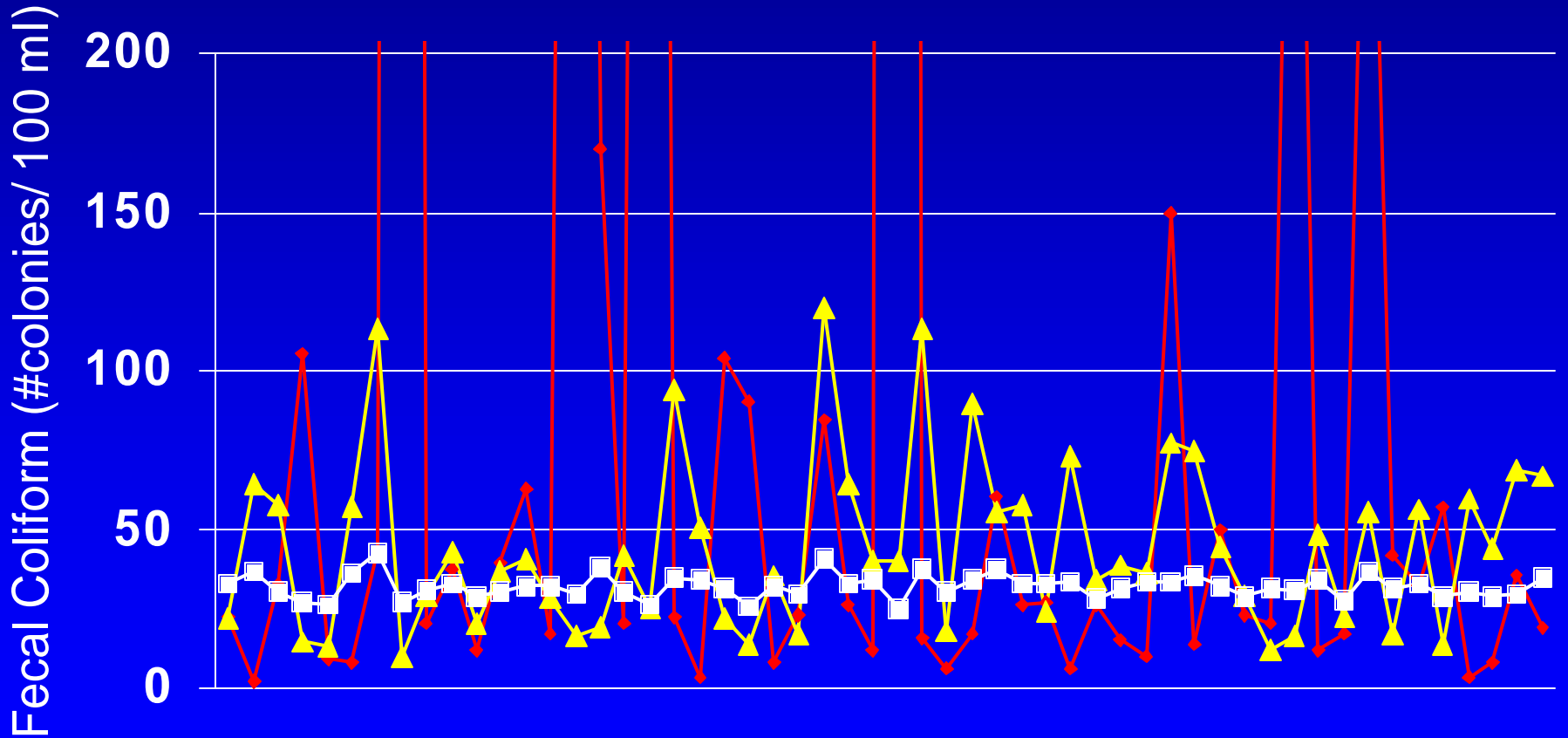
Model Performance Evaluation

Transformed Data



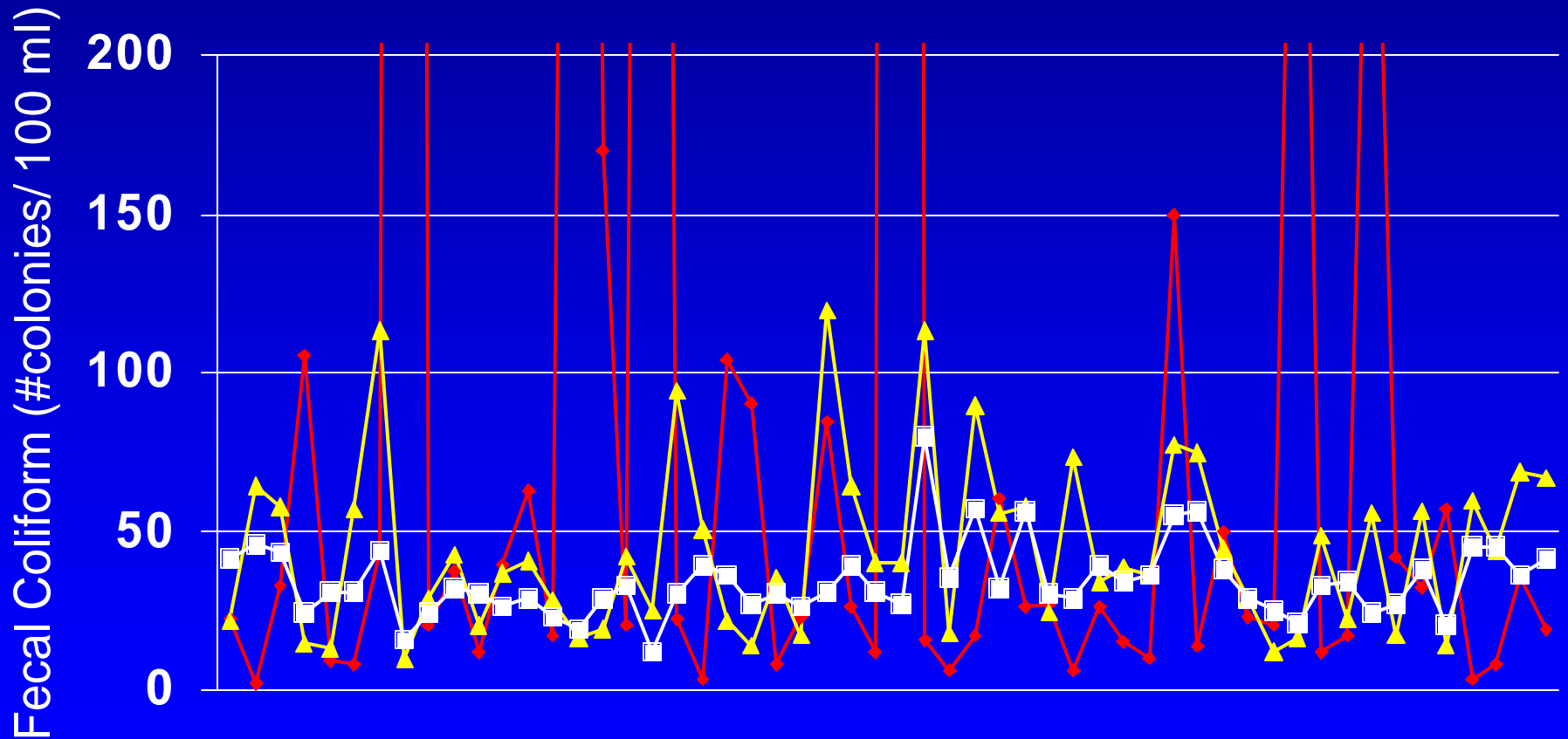
Sensitivity

—◆— Target —▲— All —■— C, S, P



Sensitivity

—◆— Target —▲— All —■— C, S, P, Tavg



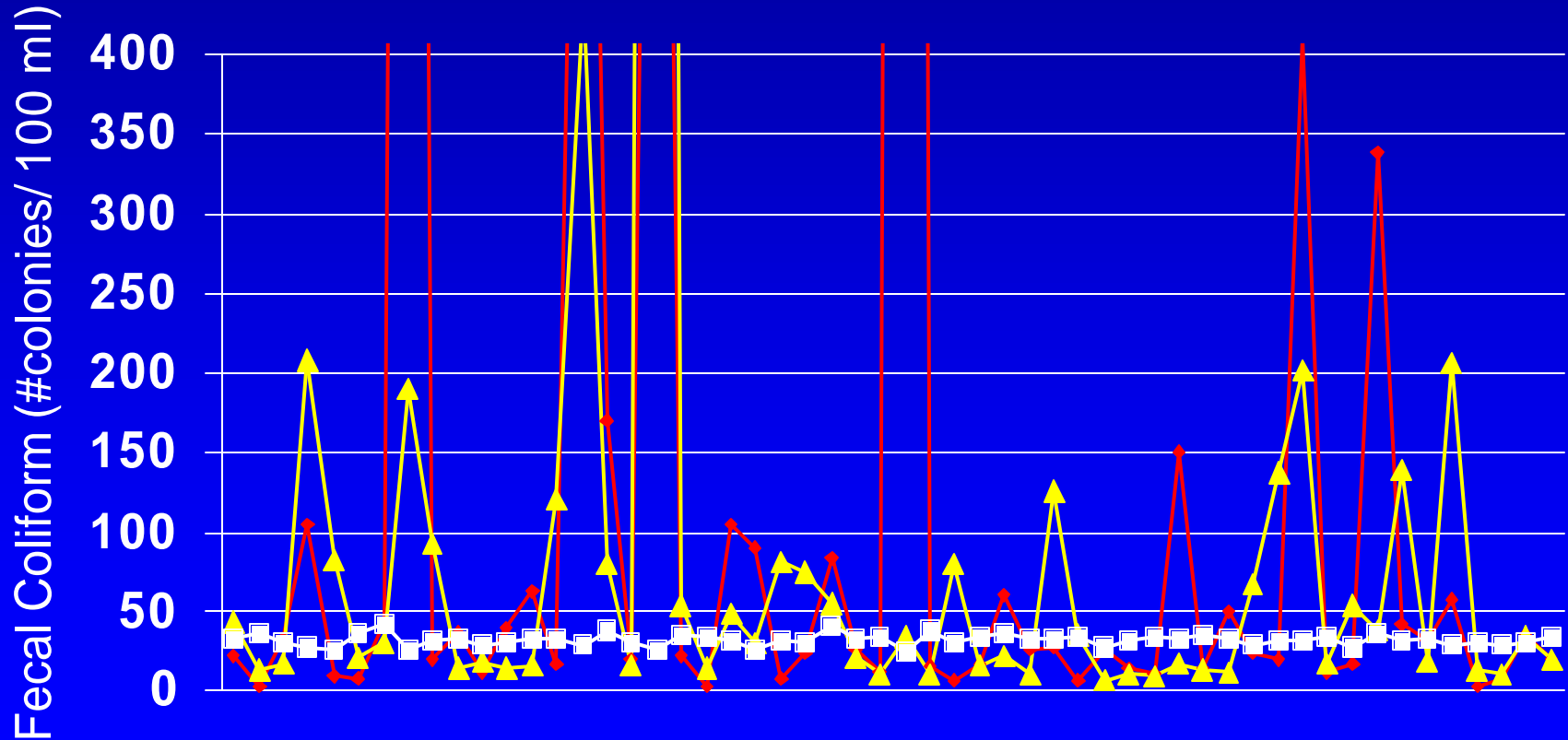
Linear Regression

- Using only parameters with significant statistical ($p < 0.05$) correlation with fecal coliform
- $\text{LN}(\text{Fecal Coliform}) = 12.00$
 - 1.17*LN(Conductivity)
 - +0.437*LN(Streamflow)
 - +0.309*LN(Precipitation)
- R-squared = 0.383 for training and validation set combined

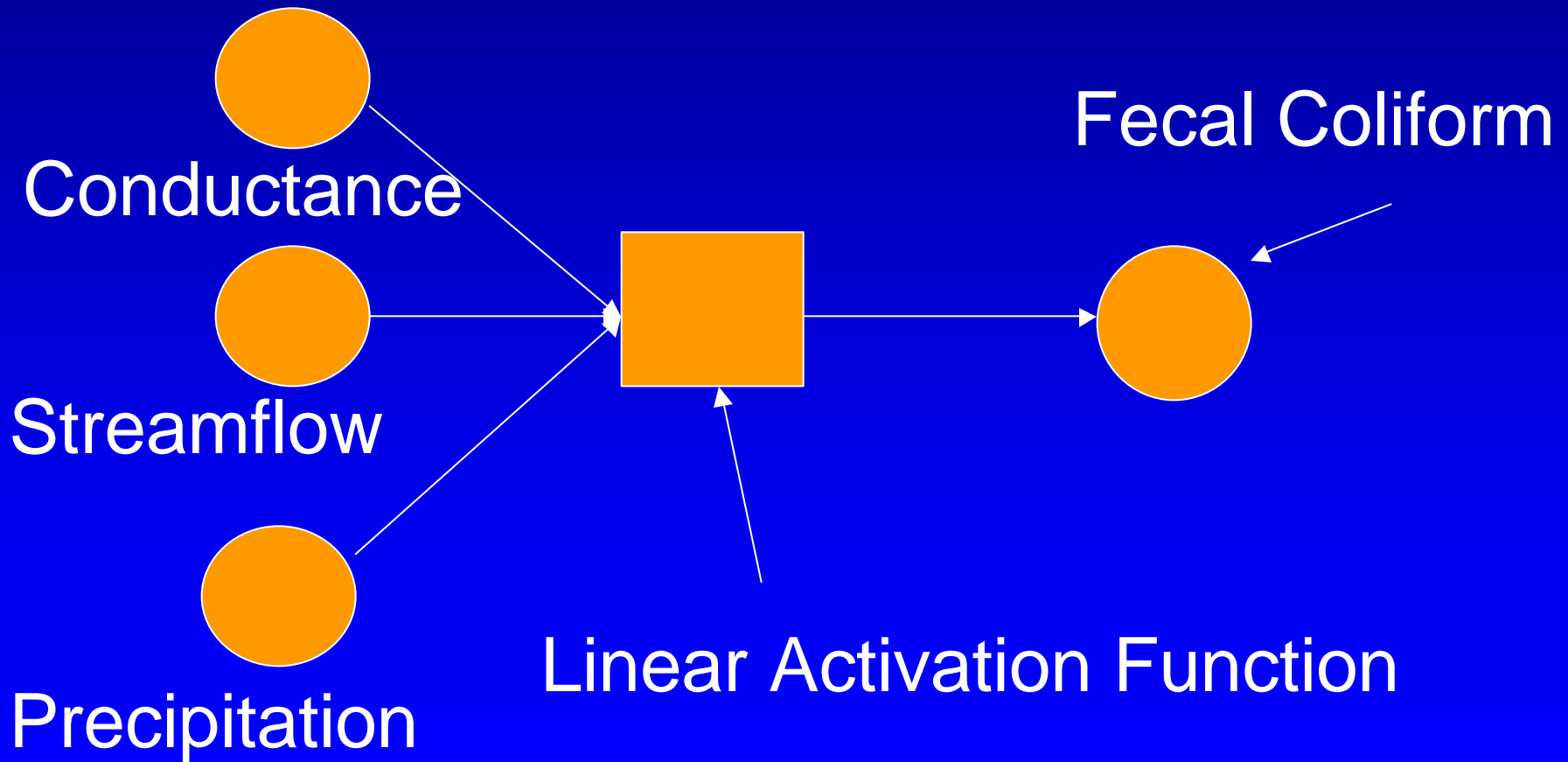
Regression
RMSE = 704
AAPE = 185

ANN
RMSE = 749
AAPE = 162

Comparison with Regression - C, S, P only



ANN as Linear Regression

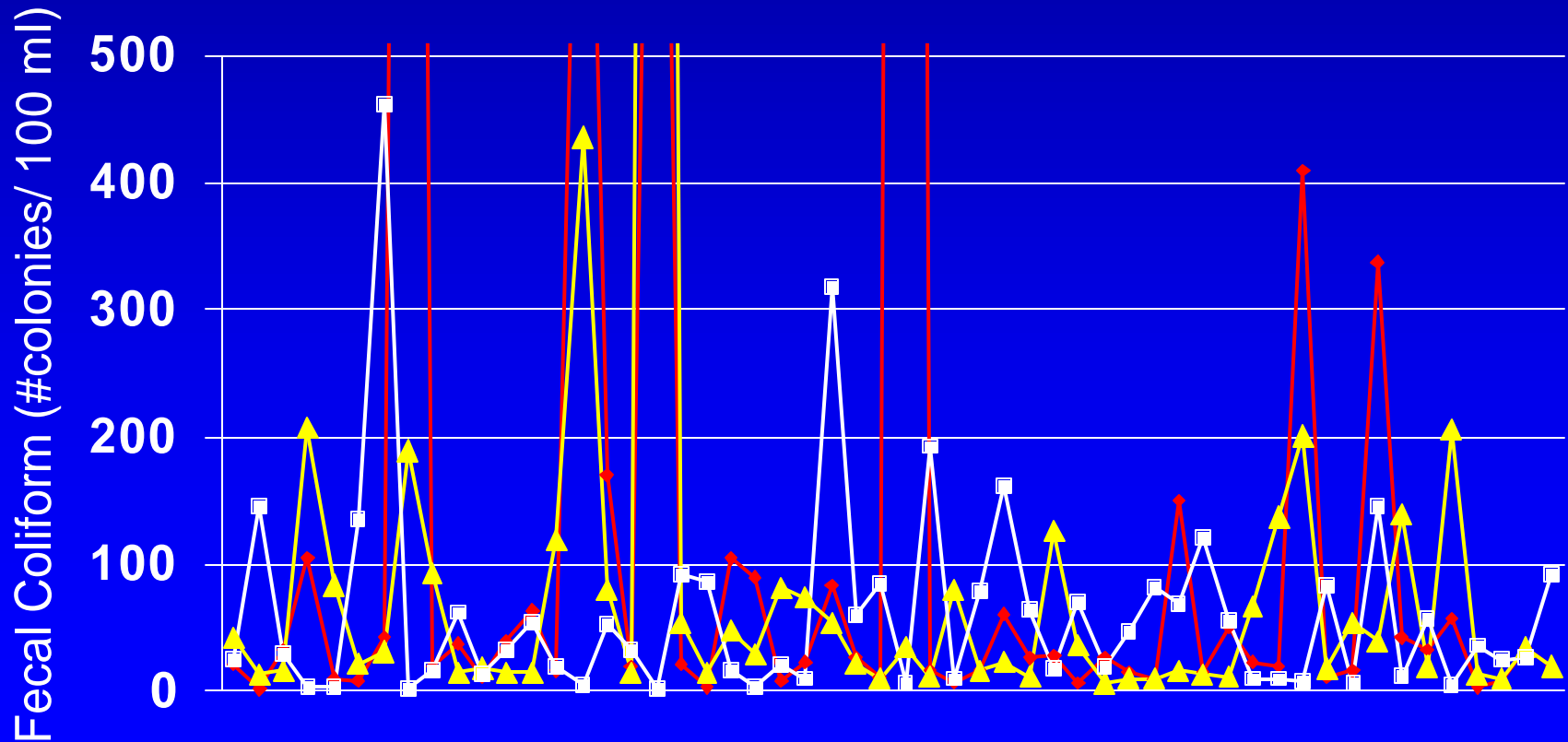


Regression
RMSE = 699
AAPE = 1612

ANN
RMSE = 1314
AAPE = 421

Comparison with Regression - C, S, P only

—●— Target —▲— Regression —□— ANN - Linear Regression



Phase I Conclusions

- Normally Distributed input improves ANN performance
- Range of Training Data is important for overall performance - reduction in RMSE and better visual comparison
- ANN has difficulty with extreme values - different processes?

Phase I Conclusions

- Using only parameters with statistical correlation to target doesn't show performance improvement
- Regression produces similar performance statistics, but better visual comparison
- Structuring ANN architecture to mimic linear regression captures variability better

Phase I Ongoing Research

- Input
 - Effect of lagged (and lumped for precipitation) input data
 - Investigation of ANN performance above a threshold fecal coliform value - different processes?
- Architecture
 - Different layer and node structures
 - Different activation functions
 - Continue comparison with statistical models

Use of Artificial Neural Networks for Modeling Indicator Organisms in a Drinking Water Supply Watershed

Diane M.L. Mas

www-unix.ecs.umass.edu/~dmas