Module 11: Time Series Modeling and Prediction of Environmental Data - Student Handout



### Name:

### Student ID:

#### Completed on:

#### Copy-paste your save progress link from the Shiny app here for ease of reference:

# **Macrosystems EDDIE Module 11: Time Series Modeling and Prediction of Environmental Data**

# Learning Objectives:

In the course of this module, you will:

* Explore relationships between ecological variables from environmental case studies.
* Understand the structure of four time series models (including machine learning models) that are commonly applied in environmental science.
* Fit time series models using environmental data and assess the importance of ecological driver variables in making model predictions.
* Process environmental datasets into standardized formats suitable for training and assessing time series models.
* Compare multiple time series models to assess their performance on out-of-sample predictions.

# Summary:

Advances in environmental sensor technology in recent decades permit collection of environmental data at high temporal frequencies (e.g., every 10 minutes) across many ecosystems. These time series of environmental data can be used to gain information about the previous and current state of ecosystems, as well as make predictions about ecosystem states in the future. To help researchers and managers understand the complex patterns that can occur in high-frequency time series data, time series models are commonly applied in ecology and environmental science. These models use statistical and machine learning methods to identify signals in high-frequency environmental data.

In this module, students will apply several different time series and machine learning models to environmental data. They will explore data from an environmental case study of their choice, choose which environmental variables to use to fit a time series model, assess the model, and use it to make predictions. Then, they will process a new dataset into a standardized format, upload it into the module, and fit several other models to compare predictive performance across models. They will also evaluate the ecological implications that can be drawn from each model (e.g., which variables are important for explaining the dynamics of the target variable for prediction).

## Module overview:

* Introduce concepts related to time series modeling and environmental data case studies
* Activity A: Visualize data from a selected environmental data case study and fit and assess a time series model
* Activity B: Choose a new environmental dataset or upload your own dataset and fit and assess a time series model
* Activity C: Fit additional time series models to your environmental dataset and compare performance across models

## Today’s focal question: *How can we use time series models to understand and predict the state of ecosystems?*

To address this question, we will introduce several time series models that are commonly applied in ecology and environmental science. We will also introduce the importance of using standardized datasets and out-of-sample prediction to train and assess time series models. We will fit a model that predicts a target variable that is important for ecosystem management. We will make out-of-sample predictions with this model to assess model fit, also examine the uncertainty associated with our predictions. Then, we will format and upload a standardized dataset and fit several models to generate out-of-sample predictions for an environmental case study that is relevant to your coursework. Finally, we will compare performance across models and draw inference regarding which ecological variables are important for predicting the target variable of management concern.

We will be using ecological data collected by the Swan-Canning Estuary Virtual Observatory (SCEVO) in Western Australia and the National Ecological Observation Network (NEON) in the United STates to tackle this question. SCEVO is a monitoring network to observe water quality in the Swan and Canning Rivers in Perth, Australia, which are important drinking water supplies. NEON is a continental-scale observatory designed to collect publicly-available, long-term ecological data to monitor changing ecosystems across the U.S.

Our focal target variable for the SCEVO dataset will be chlorophyll-a, a measure of aquatic primary productivity. Chlorophyll-a is an important ecosystem variable because it can indicate when a harmful algal bloom is occurring. High levels of chlorophyll-a may indicate a bloom, which can have deleterious effects on water quality through release of toxins, formation of surface scums, and depletion of dissolved oxygen in the water.

The target variable for the NEON dataset will be net ecosystem exchange (NEE), a measure of carbon dioxide flux from ecosystems. NEE indicates whether an ecosystem is emitting or taking up greenhouse gasses, and how this changes over time.

## R Shiny App:

The lesson content is hosted on an R Shiny App at <https://macrosystemseddie.shinyapps.io/module11/>  
This can be accessed via any internet browser and allows you to navigate through the lesson via this app. You will fill in the questions below on this handout as you complete the lesson activities.

## Optional pre-class readings and videos:

* Tredennick, A. T., Hooker, G., Ellner, S. P., & Adler, P. B. (2021). A practical guide to selecting models for exploration, inference, and prediction in ecology. Ecology, 102(6), e03336. <https://doi.org/10.1002/ecy.3336>
* Wickham, H., Çetinkaya-Rundel, M., and Grolemund, G. (2023) R for Data Science (2nd ed.), [Chapter 5: Data Tidying.](https://r4ds.hadley.nz/data-tidy.html)

## Pre-class activity: Explore environmental case study

Choose one of the papers describing an environmental case study and use it to answer the questions below. Choose the case study that you will work with for Activity A of the module.

**Phytoplankton dynamics in the Canning River, Western Australia** Huang, P., Trayler, K., Wang, B., Saeed, A., Oldham, C. E., Busch, B., & Hipsey, M. R. (2019). An integrated modelling system for water quality forecasting in an urban eutrophic estuary: The Swan-Canning Estuary virtual observatory. Journal of Marine Systems, 199, 103218. <https://doi.org/10.1016/j.jmarsys.2019.103218>

**Net ecosystem exchange in Bartlett Experimental Forest, NH, USA** Jung, Chang Gyo, and Oleksandra Hararuk. “Assimilation of NEON observations into a process‐based carbon cycle model reveals divergent mechanisms of carbon dynamics in temperate deciduous forests.” Journal of Geophysical Research: Biogeosciences 127.3 (2022): e2021JG006474. <https://doi.org/10.1029/2021JG006474>

#### Pre-class questions: Choose one of the environmental case studies above and use the paper to answer the questions below.

A. Which case study did you select?

**Answer:**

B. What ecological variable(s) are being predicted?

**Answer:**

C. How can these predictions help the public and/or managers?

**Answer:**

D. Describe the way(s) in which the model predictions are assessed (evaluated for accuracy).

**Answer:**

# Introduction

Now navigate to the [Shiny interface](https://macrosystemseddie.shinyapps.io/module11) to answer the rest of the questions.

The questions you must answer are written both in the Shiny interface as well as in this handout. As you go, you should fill out your answers in this document.

## Think about it!

Answer the following questions:

1. Why is close attention to data standards needed when using time series models in environmental science? What are the risks of not following data standards?  
     
   **Answer:**
2. How can use of time series models improve both natural resource management and ecological understanding?  
     
   **Answer:**

# Activity A: Select an environmental case study, visualize data, and fit a model

## Objective 1: Select case study

Select a case study from the table, then read the text in the ‘About Case Study’ section to find out more about the case study.

1. Fill out information about your selected case study site:  
     
   *Table 1. Site Characteristics*

| Characteristic | Answer |
| --- | --- |
| Name of selected site |  |
| Four letter site identifier |  |
| Latitude |  |
| Longitude |  |
| Elevation (m) |  |

1. What is the desired target variable for prediction in the case study you have chosen? Why is that variable important for ecosystem function and management?  
   **Answer:**

## Objective 2: Explore data

Explore the data measured at the selected site, including how each variable changes over time and how the variables relate to each other.

1. Fill out the table in your Word document with the description of site variables:  
   *Table 2. Description of site variables:*

| Variable | Units | Mean | Minimum | Maximum |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

1. Describe the relationship between each variable and the target variable. For each variable, write in the table in your Word document whether increases in that variable are associated with increases in the target variable (positive relationship), decreases in the target variable (negative relationship), or no change in the target variable (no relationship).  
   *Table 3. Description of effect of each variable on the target variable*

| Target | Variable | Relationship |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

## Objective 3: Learn about ARIMA model

We will use the environmental data from your case study to fit an ARIMA model that will predict future values of the target variable. But first, we will learn about how ARIMA models are used for time series prediction.

1. Describe the three components of an ARIMA model (autoregressive, integrated, moving average) in your own words.  
   **Answer:**
2. Observe the plot on the right. Does the target variable exhibit autocorrelation at a 1-day lag? Explain how you know.  
   **Answer:**

**Please download the plot associated with Q8 and copy-paste it into your report below.**

*Figure 1. Target variable vs. one-day lag of environmental case study target variable*

1. Observe the PACF plot on the right. Based on this figure, how many lags of the target variable do you anticipate being included in the fitted ARIMA model? Explain your reasoning.  
   **Answer:**

**Please download the plot associated with Q9 and copy-paste it into your report below.**

*Figure 2. Partial autocorrelation of environmental case study target variable*

1. Observe the plot on the right. Based on visual inspection, does differencing substantially improve the stationarity of the target variable data? Explain your reasoning.  
   **Answer:**

**Please download the plot associated with Q10 and copy-paste it into your report below.**

*Figure 3. Differenced time series of environmental case study target variable*

## Objective 4: Fit model

Fit a ARIMA time series model to the data from your chosen environmental case study.

1. List up to three exogenous regressors that you would like to include in your fitted ARIMA model. For each regressor, explain why you chose to include it in the model. If you choose not to include any regressors, explain your reasoning.  
   **Answer:**
2. Examine the (p, d, q) order of your ARIMA model printed above. Interpret the order of your ARIMA.  
     
   a. Record the order of your ARIMA model printed above.:  
     
   **Answer:**  
   b. How many lags are included in the model? How do you know?:  
     
   **Answer:**  
   c. Are the data differenced to fit the model? How do you know?:  
     
   **Answer:**  
   d. How many timesteps of previous errors are included in the model? How do you know?:  
     
   **Answer:**
3. Which term is most important in your fitted model (i.e., has the greatest absolute value)?  
   **Answer:**
4. If the standard error of a term is bigger than the absolute value of the estimate of a term, we cannot say for sure that the effect of that term is meaningfully different from 0. Are there any terms in your model that are not meaningfully different from 0? Which ones?  
   **Answer:**
5. Assess the importance of the exogenous regressors you chose to include in the model. Based on the term estimates and standard errors, were these variables important in your final, fitted model? Explain your reasoning.  
   **Answer:**

## Objective 5: Assess model fit

Assess the fit of the ARIMA model using out-of-sample test data from your chosen environmental case study.

1. Inspect the plot on the right. How well do you think the model predictions fit the observations in the test dataset? Explain your reasoning.  
   **Answer:**

**Please download the plot associated with Q16 and copy-paste it into your report below.**

*Figure 4. ARIMA predictions of out-of-sample observations of the environmental case study target variable*

1. Briefly describe in your own words how a 95% predictive interval can be estimated using the standard deviation of the model residuals.  
   **Answer:**
2. Record the standard deviation of the model residuals.  
   **Answer:**
3. Inspect the plot of model predictions with uncertainty. Now that you can see the uncertainty associated with model predictions, how well do you think the model predictions fit the observations in the test dataset? Explain your reasoning.  
   **Answer:**

**Please download the plot associated with Q19 and copy-paste it into your report below.**

*Figure 5. ARIMA predictions with uncertainty of out-of-sample observations of the environmental case study target variable*

1. Compare your answers to Q16 and Q19. Were there any differences in your interpretation of model fit when you visualized predictions with and without uncertainty? Why do you think this might be?  
   **Answer:**

# Activity B: Upload your own data and fit a time series model

## Objective 6: Upload standardized data

In this objective, you will learn about the importance of data standards for reproducible scientific analyses, as well as wrangle and upload a new environmental dataset of your choosing.

1. Describe one example of a benefit of using data standards in ecology and environmental science, as well as one example of a risk if data standards are not followed.  
     
   **Answer:**
2. Is the standardized data format for this module a long or a wide format? Explain how you know.  
     
   **Answer:**
3. Fill out the following information:  
   1. Minimum number of timesteps required for dataset  
        
      **Answer:**
   2. Maximum number of allowed unique values for ‘site\_id’.  
        
      **Answer:**
   3. Maximum number of allowed unique values for ‘variable’.  
        
      **Answer:**
   4. Required format of ‘date’ column, using Y for year, M for month, and D for day.  
        
      **Answer:**

## Objective 7: Fit an ARIMA model to your standardized dataset

Fit a ARIMA time series model to the data you uploaded in Objective 6.

1. What is the target variable for your dataset?  
     
   **Answer:**
2. List the three exogenous regressors that you would like to include in your ARIMA model. For each regressor, explain why you chose to include it in the model. If you choose not to include any regressors, explain your reasoning.  
     
   **Answer:**
3. Record the proportion of data you chose to use for model training.  
     
   **Answer:**
4. Examine the (p, d, q) order of your ARIMA model printed above. Interpret the order of your ARIMA.  
   1. Record the order of your ARIMA model printed above.  
        
      **Answer:**
   2. How many lags are included in the model? How do you know?  
        
      **Answer:**
   3. Are the data differenced to fit the model? How do you know?  
        
      **Answer:**
   4. How many timesteps of previous errors are included in the model? How do you know?  
        
      **Answer:**
5. Which term is most important in your fitted model (i.e., has the greatest absolute value)?  
     
   **Answer:**

## Objective 8: Assess model fit

Assess the fit of the ARIMA model from Objective 7 using out-of-sample test data from the dataset you uploaded in Objective 6.

1. Inspect the plot on the right. How well do you think the model predictions fit the observations in the test dataset? Explain your reasoning.  
     
   **Answer:**

**Please download the plot associated with Q32 and copy-paste it into your report below.**

*Figure 6. ARIMA predictions with uncertainty of out-of-sample observations of the standardized dataset target variable*

1. Record the RMSE and ignorance score for your model.  
     
   **Answer:**

# Activity C: Compare predictive performance across models

## Objective 9: Fit additional models

Fit three additional models to the data you uploaded in Objective 6. You will fit an autoregressive neural network model, a persistence model, and a day-of-year model.

1. Briefly describe, in your own words, each of the three additional models you will be fitting to your data.  
   1. autoregressive neural network model (NNETAR)  
        
      **Answer:**
   2. historical mean model  
        
      **Answer:**
   3. day-of-year model (DOY)  
        
      **Answer:**
2. Why are baseline or “null” models useful when conducting model comparisons?  
     
   **Answer:**
3. How many lags are used as inputs in your fitted NNETAR model, and how many neurons are in the hidden layer? Explain how you know.  
     
   **Answer:**
4. Use the interactive plot legend to examine the fit of the mean model and the DOY model. Typically, the fit of the DOY model is a curvy line, while the mean model is a flat line. Use your understanding of the structure of these two models to explain why.  
     
   **Answer:**

**Please download the plot associated with Q37 and copy-paste it into your report below.**

*Figure 7. ARIMA, NNETAR, persistence, and DOY model fits for the standardized dataset target variable*

## Objective 10: Compare model performance

Compare predictive performance across models on the testing data from the dataset you uploaded in Objective 6.

1. Which of the models has the smallest standard deviation of the residuals? Which has the largest?  
     
   **Answer:**

**Please download the plot associated with Q38 and copy-paste it into your report below.**

*Figure 8. Distributions of the residuals of ARIMA, NNETAR, persistence, and DOY model fits for the standardized dataset target variable*

1. Interpret the values of the standard deviation of the residuals for the four models. Which model do you expect to have the largest predictve uncertainty interval when you generate predictions with uncertainty? Explain your reasoning.  
     
   **Answer:**
2. Which of the models has the largest predictive uncertainty interval? Does this match what you predicted in Q39?  
     
   **Answer:**

**Please download the plot associated with Q40 and copy-paste it into your report below.**

*Figure 9. Predictions with uncertainty of ARIMA, NNETAR, persistence, and DOY models fit to the standardized dataset target variable*

1. Based on visual inspection of the plot of predictions with uncertainty, which model do you think has the highest predictive skill? Explain your reasoning.  
     
   **Answer:**
2. Record the RMSE and ignorance scores of your four models.  
     
   *Table 4. RMSE and ignorance score of the ARIMA, NNETAR, persistence, and DOY models*

| Model | RMSE | ignorance |
| --- | --- | --- |
| ARIMA |  |  |
| NNETAR |  |  |
| persistence |  |  |
| DOY |  |  |

1. Which of your models performs best according to RMSE?  
     
   **Answer:**
2. Which of your models performs best according to the ignorance score?  
     
   **Answer:**
3. Did the more complex models (ARIMA, NNETAR) perform better than the historical mean and/or DOY models? What does this result indicate about the importance of exogenous regressors for predicting the target variable in your chosen dataset?  
     
   **Answer:**

This module was developed by Lofton, M.E., M.H. Hipsey, K. Kurucz, and C.C. Carey. 31 July 2025. Macrosystems EDDIE: Time Series Modeling and Prediction of Environmental Data. Macrosystems EDDIE Module 11, Version 1. <http://module11.macrosystemseddie.org>. Module development was supported by NSF grants EF 1702506, 1926050, 2318861, DEB 2213550, and OISE 2330211.

This document last modified: 31 July 2025 by MEL.