### JOURNAL PUBLICATION CITATION: Whitney M. Woelmer, R. Quinn Thomas, Mary E. Lofton, Ryan P. McClure, and Cayelan C. Carey. 2021. Near-term iterative forecasts of phytoplankton at multiple time scales demonstrate the effect of model time step and forecast horizon on predictability. *Ecological Applications.*

### Data S1

### Historical model selection repository

### Author(s) [of the material provided in DataS1.zip]

Whitney Woelmer  
Department of Biological Sciences, Virginia Tech

926 W. Campus Dr.

Blacksburg, VA, USA  
wwoelmer@vt.edu

R. Quinn Thomas  
Department of Forecast Resources and Environmental Conservation  
310 W. Campus Dr.

Blacksburg, VA, USA  
rqthomas@vt.edu

Mary E. Lofton  
Department of Biological Sciences, Virginia Tech

926 W. Campus Dr.

Blacksburg, VA, USA  
melofton@vt.edu

Ryan P. McClure  
Department of Biological Sciences, Virginia Tech

926 W. Campus Dr.

Blacksburg, VA, USA  
ryan333@vt.edu

Cayelan C. Carey  
Department of Biological Sciences, Virginia Tech

926 W. Campus Dr.

Blacksburg, VA, USA  
cayelan@vt.edu

### File list (files found within DataS1.zip)

Folders

**Description**

**ARIMA\_data**: folder for storing data files created in this repository

**Correlation matrices**

**Selection Comments.xlsx:** Microsoft Word document detailing the method for removing correlated driver variables

**chlalog\_1.0m:** folder containing correlation matrices for logged chl data at 1.0m

**correlation\_matrix\_chlalog\_2013\_2016.csv:** correlation matrix of logged chl-a and driver data from 2013-2016

**correlation\_matrix\_chlalog\_2013.csv:** correlation matrix of logged chl-a and driver data from 2013

**correlation\_matrix\_chlalog\_2014.csv:** correlation matrix of logged chl-a and driver data from 2014

**correlation\_matrix\_chlalog\_2015.csv:** correlation matrix of logged chl-a and driver data from 2015

**correlation\_matrix\_chlalog2016.csv:** correlation matrix of logged chl-a and driver data from 2016

**chlasqrt\_1.0m:** folder containing correlation matrices for logged chl data at 1.0m

**correlation\_matrix\_chlasqrt\_2013\_2016.csv:** correlation matrix of square-root transformed chl-a and driver data from 2013-2016

**correlation\_matrix\_chlasqrt\_2013.csv:** correlation matrix of square-root transformed chl-a and driver data from 2013

**correlation\_matrix\_chlasqrt\_2015.csv:** correlation matrix of square-root transformed chl-a and driver data from 2015

**correlation\_matrix\_chlasqrt2016.csv:** correlation matrix of square-root transformed chl-a and driver data from 2016

**correlation\_matrix\_chlasqrt\_2013\_2016\_selected.csv:** correlation matrix of square-root transformed chl-a and selected driver data from 2013-2016 after removing driver variables which are highly correlated (this was done in Microsoft Xcel)

**correlation\_matrix\_chlasqrt\_2013\_selected.csv:** correlation matrix of square-root transformed chl-a and selected driver data from 2013 after removing driver variables which are highly correlated (this was done in Microsoft Xcel)

**correlation\_matrix\_chlasqrt\_2014\_selected.csv:** correlation matrix of square-root transformed chl-a and selected driver data from 2014 after removing driver variables which are highly correlated (this was done in Microsoft Xcel)

**correlation\_matrix\_chlasqrt\_2015\_selected.csv:** correlation matrix of square-root transformed chl-a and selected driver data from 2015 after removing driver variables which are highly correlated (this was done in Microsoft Xcel)

**correlation\_matrix\_chlasqrt\_2016\_selected.csv:** correlation matrix of square-root transformed chl-a and selected driver data from 2016 after removing driver variables which are highly correlated (this was done in Microsoft Xcel)

**correlation\_matrix\_chlasqrt\_2013\_2016\_selected\_predictable.csv:** correlation matrix of square-root transformed chl-a and selected driver data from 2013-2016 after first removing any driver variables which are not readily predictable as forecasted driver data, and second which are highly correlated with each other (this was done in Microsoft Xcel)

**correlation\_matrix\_chlasqrt\_2013\_selected\_predictable.csv:** correlation matrix of square-root transformed chl-a and selected driver data from 2013 after first removing any driver variables which are not readily predictable as forecasted driver data, and second which are highly correlated with each other (this was done in Microsoft Xcel)

**correlation\_matrix\_chlasqrt\_2013\_2016\_fortnightly\_data.R:** correlation matrix of square-root transformed chl-a and driver data from 2013-2016 on a fortnightly timestep

**selected\_correlation\_matrix\_chlasqrt\_2013\_2016\_fortnightly\_data.R:** correlation matrix of square-root transformed chl-a and selected driver data from 2013-2016 on a fortnightly timestep after first removing any driver variables which are not readily predictable as forecasted driver data, and second which are highly correlated with each other (this was done in Microsoft Xcel)

**dailymodel\_correlation\_matrix\_chlasqrt\_2013\_2016.R:** correlation matrix of square-root transformed chl-a and driver data from 2013-2016 on a daily timestep

**dailymodel\_correlation\_matrix\_chlasqrt\_2013\_2016\_selectedvariables.R:** correlation matrix of square-root transformed chl-a and selected driver data from 2013-2016 on a daily timestep after first removing any driver variables which are not readily predictable as forecasted driver data, and second which are highly correlated with each other (this was done in Microsoft Xcel)

**Interpolations**

**ctd\_interp.csv:** file of CTD data on a weekly timestep

with any missing data interpolated using na.approx()

**TP\_interp.csv:** data file containing total phosphorous data with any missing weeks interpolated

**TN\_interp.csv:** data file containing total nitrogen data with any missing weeks interpolated

**TPinf\_interp.csv:** data file containing total phosphorous data from the inflow to the reservoir with any missing weeks interpolated

**TNinf\_interp.csv:** data file containing total nitrogen data from the inflow to the reservoir with any missing weeks interpolated

**nutsother\_interp.csv:** data file containing other nutrient data with any missing weeks interpolated

**doc\_interp.csv:** data file containing dissolved organic carbon data with any missing weeks interpolated

**kd\_interp.csv:** data file containing kd, light extinction coefficient, data with any missing weeks interpolated

**collated\_interpolated.csv**: file containing all the individual interpolated datasets

**arima\_data\_fortnightly.csv:** data file containing driver data and square-root transformed chl-a data on a fortnightly timestep

**data\_arima\_WW.csv:** data file of response and driver data after model selection for autoregressive linear model

**data\_interpolated\_MayOct13\_16.csv**: file containing formatted data for CTD, chemistry, and Kd data including interpolations from 2013 to 2016 for FCR

**data\_interpolated\_plusinflowcalcs\_MayOct13\_16.csv:** all data, including interpolated data points at FCR, for CTD, chemistry, and inflow metrics (discharge, temperature, nutrient loads)

**EXO\_plusdrivers\_AR.csv:** file of transformed response and driver data on a daily timestep

**FCR\_CTD\_50\_binned.csv:** file of CTD data at Site 50 ‘binned’ into discrete depths

**FCR\_VT\_data\_2013\_2017.csv:** data file containing CTD, water chemistry, Kd, and inflow data for FCR

**Important dates FCR\_073018.csv:** file containing important dates with management implications for FCR

**interpolated\_weeks\_2013\_2016.csv:** data file containing observed and interpolated data from CTD, nutrient chemistry, and inflow datasets, as well as Julian week identifiers

**variables\_all\_2013\_2016.csv:** data file containing all driver variables from 2013 to 2016

**variables\_all\_pluslag\_2013\_2016.csv:** data file containing all driver variables plus a 1-week lag of chl-a from 2013-2016

**model\_transformed\_chlalog\_2013\_2016.csv:** data file containing log-transformed chl-a data

**model\_transformed\_chlasqrt\_2013\_2016.csv:**data file containing square root-transformed chl-a data

**Data:**folder for storing original data files which contribute to analyses in this repository

**CTD**

**CTD\_FCR.csv:** CTD data from FCR at multiple sites and depths ranging from 2013-2017. Data are now published atdoi.org/10.6073/pasta/5448f9d415fd09e0090a46b9d4020ccc

**FCR\_CTD\_1m\_2018.csv:** 2018 CTD data from FCR at 1m. Data are now published atdoi.org/10.6073/pasta/5448f9d415fd09e0090a46b9d4020ccc

**CTD\_notmatlab\_ready\_2019\_fcr50.csv:** data file of observed CTD data at FCR site 50 (deep hole) from 2019. Data are now published atdoi.org/10.6073/pasta/5448f9d415fd09e0090a46b9d4020ccc

**FCR\_CTD\_50\_1m\_2018.csv:** data file of observed CTD data at FCR site 50 (deep hole) at 1.0m depth in 2018. Data are now published atdoi.org/10.6073/pasta/5448f9d415fd09e0090a46b9d4020ccc

**FCR\_CTD\_2018\_2019.csv:** data file of observed CTD data at FCR from 2018-2019. Data are now published atdoi.org/10.6073/pasta/5448f9d415fd09e0090a46b9d4020ccc

**FCR\_CTD\_2019\_50\_binned.csv:** data file of observed CTD data at FCR site 50 (deep hole) from 2019 binned into distinct depths

**Fluoroprobe**

**FCR2014\_FluoroMeta.csv:** data file of fluoroprobe data from 2014

**FCR2015\_FluoroMeta.csv:** data file of fluoroprobe data from 2015

**FCR2016\_FluoroMeta50.csv:** data file of fluoroprobe data from 2016 from Site 50 at FCR

**Fluoro\_FCR50\_2014\_2017.csv:** data file of fluoroprobe data from 2014 to 2017 created in script 4

**FC\_recal\_2017.txt:** data file of 2017 Fluoroprobe data

**Inflow**

**FCR\_inflow\_WVWA\_2013\_2019.csv:** data file containing observed pressure and temperature data, including discharge calculations from the main inflow to FCR. This file can be accessed from github (<https://github.com/CareyLabVT/Reservoirs/blob/master/Data/DataAlreadyUploadedToEDI/EDIProductionFiles/MakeEMLInflow/2019/inflow_for_EDI_2013_2019.csv>) and is published as a previous version of this data product on EDI (doi.org/[10.6073/pasta/8d22a432aac5560b0f45aa1b21ae4746](http://resolver.ebscohost.com/openurl?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&__char_set=utf8&rft_id=info:doi/10.6073/pasta/8d22a432aac5560b0f45aa1b21ae4746&rfr_id=info:sid/libx%3Avirginiatech&rft.genre=article))

**FCRweir.csv:** this file contains observed pressure and temperature data from a pressure transducer owned by Virginia Tech at the main inflow to FCR which was deployed in 2018. This file can be downloaded from github (https://github.com/FLARE-forecast/FCRE-data/tree/fcre-weir-data), and is now part of a published data package on EDI (doi.org/[10.6073/pasta/8d22a432aac5560b0f45aa1b21ae4746](http://resolver.ebscohost.com/openurl?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&__char_set=utf8&rft_id=info:doi/10.6073/pasta/8d22a432aac5560b0f45aa1b21ae4746&rfr_id=info:sid/libx%3Avirginiatech&rft.genre=article))

**inflowcalcs\_FCR.csv:** data file containing inflow data for FCR (mean water residence time and discharge, daily temperature and discharge summary statistics), calculated using the EDI dataset, FCR\_inflow\_WVWA\_2013\_2019.csv

**MET**

**FCR\_GLM\_met\_NLDAS2\_010113\_010118\_GMTadjusted.csv:** meteorological data downloaded from the National Land and Data and Assimilation System (NLDAS-2; Xia et al. 2012) for FCR from 2013 to 2018

**FCR\_GLM\_met\_NLDAS2\_Dec14\_Dec18.csv:** meteorological data downloaded from the National Land and Data and Assimilation System (NLDAS-2; Xia et al. 2012) for FCR from 2014 to 2018

**FCR\_GLM\_NLDAS\_010113\_123118.csv:** NLDAS data from 2013-2018 at FCR, derived by merging together FCR\_GLM\_met\_NLDAS2\_010113\_010118\_GMTadjusted.csv and FCR\_GLM\_met\_NLDAS2\_Dec14\_Dec18.csv

**FCRmet.csv:** observed meteorological data from FCR, downloaded from github (https://github.com/FLARE-forecast/FCRE-data/tree/fcre-metstation-data), where it is stored in files organized by year due to the large file size. However, a published version of the data can be found at: doi.org/[10.6073/pasta/890e4c11f4348b3ceda802732ffa48b4](http://resolver.ebscohost.com/openurl?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&__char_set=utf8&rft_id=info:doi/10.6073/pasta/890e4c11f4348b3ceda802732ffa48b4&rfr_id=info:sid/libx%3Avirginiatech&rft.genre=article)

**FCRmet\_legacy01.csv:** observed meteorological data from FCR for 2018 data, downloaded from github (https://github.com/FLARE-forecast/FCRE-data/tree/fcre-metstation-data), where it is stored in files organized by year due to the large file size. However, a published version of the data can be found at: doi.org/[10.6073/pasta/890e4c11f4348b3ceda802732ffa48b4](http://resolver.ebscohost.com/openurl?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&__char_set=utf8&rft_id=info:doi/10.6073/pasta/890e4c11f4348b3ceda802732ffa48b4&rfr_id=info:sid/libx%3Avirginiatech&rft.genre=article)

**Met\_FCR\_daily.csv:** daily estimates of meteorological variables at FCR from 2013-2018. Data derived from NLDAS dataset in script 2\_summarize\_daily\_met\_TIDY.r

**NLDAS\_2018\_daily.csv:** NLDAS data on a daily timestep from 2018 at FCR, derived by merging together FCR\_GLM\_met\_NLDAS2\_010113\_010118\_GMTadjusted.csv and FCR\_GLM\_met\_NLDAS2\_Dec14\_Dec18.csv

**update\_met.csv:** this is a working file that is produced as an output from the function create\_obs\_met\_input()

**update\_met.csv:** this is a working file that is produced from the function create\_obs\_met\_input()in Data S2

**Water\_chemistry**

**FCR\_chemistry:** nutrient chemistry data for FCR from 2013 to 2017, this is a subset of data published in a previous version of the EDI package (doi.org/[10.6073/pasta/8d83ef7ec202eca9192e3da6dd34a4e0](http://resolver.ebscohost.com/openurl?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&__char_set=utf8&rft_id=info:doi/10.6073/pasta/8d83ef7ec202eca9192e3da6dd34a4e0&rfr_id=info:sid/libx%3Avirginiatech&rft.genre=article) )

**YSI\_PAR\_SECCHI**

**YSI\_PAR\_profiles.csv:** data file of photosynthetically active radiation and secchi data for FCR collected using a YSI sonde. This is published as a previous version of the EDI data product (doi.org/[10.6073/pasta/3e9f27971e353c8a80840b5e99a67d0c](http://resolver.ebscohost.com/openurl?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&__char_set=utf8&rft_id=info:doi/10.6073/pasta/3e9f27971e353c8a80840b5e99a67d0c&rfr_id=info:sid/libx%3Avirginiatech&rft.genre=article))

**FCR\_Kd.csv:** data file of Kd, light extinction coefficient

**Rscripts:** folder for storing code for this repository

**.Rhistory**: Rhistory file for this repository

**1\_FCR\_CTDlayers\_Site50\_ONLY\_replacedepth.R**: Script to create “layers” of CTD data from high spatial frequency profile. Takes all the CTD data and selects data for specific depths (based on the nutrient chemistry sampling depths). Replace CTD depth with the sampling depth to create “layers” (e.g., 0.1m, 0.8m, etc.). Input file: CTD\_FCR.csv. Creates file: FCR\_CTD\_50surf\_binned.csv, FCR\_CTD\_50\_binned.csv

**2\_summarize\_daily\_met\_TIDY.R:** script to calculatesummary statistics (mean, median, sum, etc.) for meteorological data to go from hourly met data to daily statistics

**3\_calculate\_spcond\_CTD.R:** Script to calculate specific conductance for all CTD data at site 50 and remove CTD-generated specific conductance. Reads in file FCR\_CTD\_50\_binned.csv, creates new columns, and writes again to same file.

**4\_format\_fluora.R:**script to collate and format observed fluoroprobe phytoplankton fluorescence data from 2014-2017. Creates file: Fluoro\_FCR50\_2014\_2017.csv

**5\_calculate\_Kd.R:** script to calculate, Kd, light extinction coefficient.

**6\_merge\_FCR\_data.R:** script to merge various datasets together to create a dataframe for FCR time series on a weekly timestep.

**7\_timestep.R**: script to explore the weekly frequency of data and set a regular timestep interval. Week without observations are identified (n = 18 datapoints)

**8\_interpolate\_allcollecteddata.R**: Script to interpolate days where chla does not have data in order to reach a weekly timestep

**9\_inflow.R**: script to calculate residence time and nutrient loads for FCR based on inflow data from main inflow to FCR

**10\_weeks.R**: script to select one datapoint per week when more than one is available and calculate some weekly summary identifiers

**11\_merge\_weekly\_met.R:** script to merge together interpolated driver data and meteorological data

**12\_ARlag.R**: script to example partial autocorrelation of time lags for chl-a data and to add a new column with the appropriate lag

**13\_transformations.R**: script to assess normality of driver and response variables and transform variables according. To meet the normality assumptions for response variable (chl-a) we produce files with chl-a either log-transformed and a square root-transformed

**14\_correlation\_coefficients\_chlalog.R**: script to create correlation matrices for log-transformed chl-a data. This is done for the entire dataset (2013-2016) as well for individual years. Once matrices are produced, driver variables that are highly correlated with each other are eliminated through the process described in the script.

**14\_correlation\_coefficients\_chlasqrt.R**: script to create correlation matrices for square root-transformed chl-a data. This is done for the entire dataset (2013-2016) as well for individual years. Once matrices are produced, driver variables that are highly correlated with each other are eliminated through the process described in the script.

**14\_correlation\_coefficients\_chlasqrt\_fortnightly.R:** script to create correlation matrices for square root-transformed chl-a data on a fortnightly timestep. This is done for the entire dataset (2013-2016) as well for individual years. Once matrices are produced, driver variables that are highly correlated with each other are eliminated through the process described in the script.

**15\_lm\_iterations\_chlasqrt\_1.0m\_2013\_2016\_predictable:** script to create a global model using all selected covariates, create a dataframe with only the best performing covariates after model selection, and write that to a file

**16\_merge\_NLDAS\_years.**R: script to merge together 2013-2018 NLDAS data

**17\_2018\_Validation.**R: script to collate response and driver data from 2018, perform appropriate transformations, and make prediction using the model developed from 2013-2016 data

**18\_create2018\_2019\_CTD\_input.R**: script to collate 2018 and 2019 CTD data onto 2013-2016 dataset. Reads in observed CTD profile with high spatial resolution and bins into discrete depths

**19\_lm\_iterations\_fortnightly\_data.R**: script to run model selection on data at a fortnightly timestep to see if drivers differ from weekly model selection

**20\_lm\_iterations\_daily\_data.R**: script to collate driver and response data on a daily timestep and run full model selection, also writes a data file of chl and selected driver variables (relative humidity)

**21\_compare\_interpolated\_and\_no\_interpolated\_training\_data.R**: script to compare linear model performance of the training data set which includes interpolated datapoints (n=18) and a dataset with no interpolated datapoints

**create\_obs\_met\_input.R**: function which takes the observed met data at FCR and calculates hourly summaries for a given date range

**temp\_oxy\_chla\_qaqc.R**: function to extract observations of temperature, oxygen, and chla from observed data at FCR and perform QAQC. The script reads in both a file with raw data and a maintenance log and removes data when maintenance was being performed on the sensors. Other QAQC for when sensors experienced extreme fouling is also performed (see EDI data publication of this data product for more details: (doi:[10.6073/pasta/88896f4a7208c9b7bddcf498258edf78](http://resolver.ebscohost.com/openurl?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&__char_set=utf8&rft_id=info:doi/10.6073/pasta/88896f4a7208c9b7bddcf498258edf78&rfr_id=info:sid/libx%3Avirginiatech&rft.genre=article)). Lastly, the function writes the output to a specific location.

**Metadata\_DataS1.docx:** file documenting metadata for this repository