ArboMAP User Guide

Arbovirus Modeling and Prediction: West Nile Virus Forecasting

Dawn M. Nekorchuk, Justin K. Davis, and Michael C. Wimberly
(mcwimberly@ou.edu)
Ecological and Geospatial Reseearch and Applications in Planetary Health (EcoGRAPH)
Geography and Environmental Sustainability, University of Oklahoma

Updated May 13, 2022 for Version 4.0

Contents

1 Introduction

1.1 ArboMAP forecasting system overview

The Arbovirus Monitoring and Prediction (ArboMAP) system produces a weekly, county-level forecast of human West Nile virus (WNV) cases using environmental data combined with entomological data (Figure 1).

The transmission of mosquito-borne diseases, such as WNV, is influenced by environmental conditions that affect many aspects of the disease transmission system. ArboMAP uses an ensemble of different mathematical models that each are predicting if a county will report at least one case in a given CDC/MMWR epiweek ('positive county-week'). Results presented are an average of the models with ranges as appropriate. As part of the process, mosquito infection rate is also modeled based on the mosquito pool data, and is included in the default modeling. ArboMAP uses generalized additive models (GAMs) with smooths for seasonality, and also lagged weather data, which allows it to model the time-delayed effects of weather conditions. A forecast report appendix, if generated, will expand all the results to show each individual model.

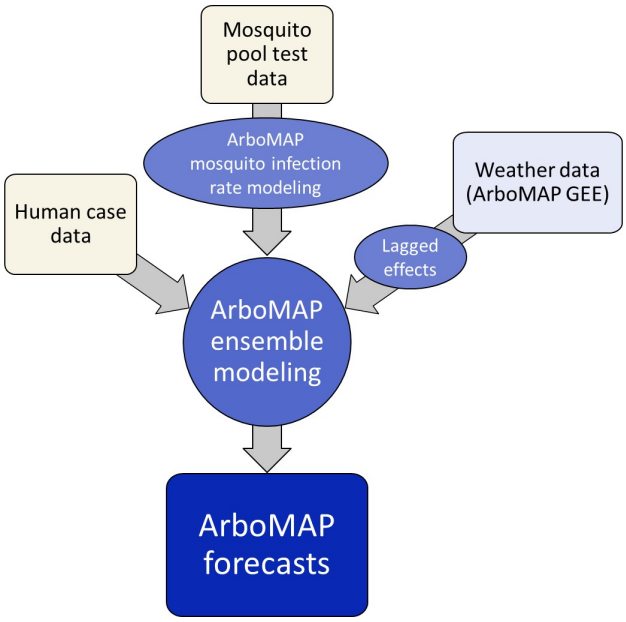

Figure 1: Overview of the data, modeling, and output flows of the ArboMAP system. Three different sources of data are given to the system to produce forecasts: human surveillance data, mosquito pool test results (entomological), and weather data. Weather data can be obtained through the ArboMAP Google Earth Engine (GEE) tool.

1.2 WNV history and modeling

West Nile virus (WNV) in South Dakota is the example in this document. WNV circulates primarily among mosquitoes and birds, but occasionally spills over into human or other hosts. In humans, the majority (~80%) of infections show no symptoms and are not diagnosed except perhaps accidentally during blood donation. Some small proportion of individuals will develop symptoms such as headache, fever, and rash, but in about one out of every 100 infected individuals, the disease becomes neuroinvasive and can cause debilitating illness and death.

WNV was first introduced to the U.S. in 1999 on the east coast, after which it spread in a westward-moving wave across the country. The early years produced thousands of cases as WNV invaded new regions. Birds

were immunologically naive, so the virus decimated entire avian populations and spilled over into humans in an outbreak that lasted several years in some states.

In South Dakota, there were more than a thousand reported cases in 2002-2003 (Wimberly et al. 2012), which implies that many more were actually infected, since most infected individuals do not show symptoms and are never diagnosed. After a number of years of active transmission, case numbers declined and it looked like the disease might vanish from circulation after 2011, in which there were only two cases reported in South Dakota.

However, in 2012 there was another outbreak, with hundreds of cases in South Dakota (Wimberly et al. 2013). Because of these fluctuations in annual case counts, there is a need to predict the magnitude and locations of WNV in advance to support proactive mosquito control and disease prevention activities. Thus, a key scientific question is whether there are there any environmental or entomological indicators that could inform, in advance, how much WNV transmission is likely to occur in the current year.

Multiple studies conducted in South Dakota have confirmed that vector mosquitoes and human WNV cases are sensitive to fluctuations in environmental variables, particularly temperature (Chuang et al. 2011, 2012a, 2012b; Hess et al. 2018; Wimberly et al. 2008, 2014). EcoGRAPH's research has shown that relatively simple statistical models, relying on weather and mosquito infection data, can be used to predict the risk of infection on a district-week basis (Davis et al. 2017, 2018). For more background, relevant journal articles have been listed at the end of this document.

1.3 ArboMAP modeling

ArboMAP uses an ensemble of different mathematical models that each are predicting if a county will report at least one case in a given week ('positive county-week'). It uses "big additive models" run with the `bam()` function from the mgcv library. Predictor variables include weather variables with effects modeled as distributed lags, and mosquito infection rates modeled using one of several techniques. There are a variety of other modeling options:

- Models can be based on 1) untransformed weather data, or 2) weather anomalies, calculated as the difference between each daily value and its long-term daily expectation, combined with a cyclical seasonal trend.
- Distributed lags can be 1) a single, fixed set of distributed lags, or 2) seasonally varying so that the lag functions can change over the course of the WNV season.
- Smoothed functions can be modeled using 1) cubic splines, or 2) thin-plate splines.

Models to be used for the forecasts are specified in a text file - in the demo, `models.txt` in the `data_models` folder of the ArboMAP project. Each line in this file contains two comma-separated strings, with **no** space between the fields. The first field contains a code name for the model and the second field contains the specification of the R model formula. In the file provided with ArboMAP, the naming convention describes whether cubic (cub) or thin-plate (tp) splines are used. Then, the distributed lags are either fixed (fx) or seasonally-varying (sv). Then, either anomalized (anom) or non-anomalized (nonanom) environmental data are used.

The table below lists the models that are found in the demo `models.txt` file. The following fields may be present:

- `any_cases`: positive county-week
- `arbo_ID`: internal field for identifying counties
- `mir_stat`: the mosquito infection rate statistic
- `s(lag, by=var...)`: fixed smooth term for the environmental variable over the distributed lag period
- `te(lag, doymat, by=var...)`: seasonally-varying smooth term for the environmental variable over the distributed lag period

- `var1`: variable for parameter `predictor_var1`, observed value
- `var2`: variable for parameter `predictor_var2`, observed value
- `var1_anom`: variable for parameter `predictor_var1`, anomalized value
- `var2_anom`: variable for parameter `predictor_var2`, anomalized value
- `s(doy,...)`: smooth term for day of year, for seasonality

Model	Description	Formula
cub-fx-nonanom	Non-anomalized weather with fixed cubic splines	any_cases ~ 0 + arbo_ID + mir_stat + s(lag, by=var1, bs='cr') + s(lag, by=var2, bs='cr')
cub-fx-anom	Anomalized weather with fixed cubic splines	any_cases ~ 0 + arbo_ID + mir_stat + s(lag, by=var1_anom, bs='cr') + s(lag, by=var2_anom, bs='cr') + s(doy, bs='cr')
cub-sv-nonanom	Non-anomalized weather with seasonally-varying cubic splines	any_cases ~ 0 + arbo_ID + mir_stat + te(lag, doymat, by=var1, bs='cr') + te(lag, doymat, by=var2, bs='cr')
cub-sv-anom	Non-anomalized weather with seasonally-varying cubic splines	any_cases ~ 0 + arbo_ID + mir_stat + te(lag, doymat, by=var1_anom, bs='cr') + te(lag, doymat, by=var2_anom, bs='cr') + s(doy, bs='cr')
tp-fx-nonanom	Non-anomalized weather with fixed thin plate splines	any_cases ~ 0 + arbo_ID + mir_stat + s(lag, by=var1, bs='tp') + s(lag, by=var2, bs='tp')
tp-fx-anom	Anomalized weather with fixed thin plate splines	any_cases ~ 0 + arbo_ID + mir_stat + s(lag, by=var1_anom, bs='tp') + s(lag, by=var2_anom, bs='tp') + s(doy, bs='tp')
tp-sv-nonanom	Non-anomalized weather with seasonally-varying thin plate splines	any_cases ~ 0 + arbo_ID + mir_stat + te(lag, doymat, by=var1, bs='tp') + te(lag, doymat, by=var2, bs='tp')
tp-sv-anom	Anomalized weather with seasonally-varying thin plate splines	any_cases ~ 0 + arbo_ID + mir_stat + te(lag, doymat, by=var1_anom, bs='tp') + te(lag, doymat, by=var2_anom, bs='tp') + s(doy, bs='tp')

In general, it is expected that models based on thin-plate splines will be more stable and perform better than models based on cubic splines. However, this may not always be the case, and fitting models based on thin-plate splines typically takes longer. Models based on weather anomalies tended to outperform those based on transformed weather variables, though again this is not necessarily always the case. Models with seasonally-varying coefficients can be more effective when environmental sensitivities are different early in the season (when virus amplification is occurring in bird populations) versus later in the season (when infected mosquitoes are biting humans). However, the seasonally varying models are more complex, take more time to fit, and may be more likely to overfit to the training dataset.

Other important modeling choices include determining which meteorological variables to use as predictors, and which type of mosquito index to use. Based on previous research, the best results seem to be obtained by combining temperature with a variable related to moisture (vapor pressure deficit, relative humidity, or precipitation). It is also necessary to choose a technique for modeling the mosquito infection data to generate entomological risk indices. In most situations, the mosquito infection growth rate (MIGR) is the best model of entomological risk, particularly in the early part of the WNV season. In situations where the mosquito infection rates does not exhibit unimodal growth in the early season, the mosquito infection intercept (MII) or the area under the mosquito infection growth curve (AUC) can be used as alternatives. For MIGR and MII, there is the also the option of implementing a spatially stratified version where multiple can be use in different parts of the state.

The report generated by ArboMAP provides fit statistics for each model formula included in the run. Only one set of meteorological variables and one mosquito infection model can be selected for each run. Therefore,

multiple runs must be made to compare different options for these settings. A suggested strategy is to start with multiple runs that include all the model formulas combined with environmental variables and mosquito infection models. Then, a smaller subset of the best-fitting models can be selected to use for routine forecasting.

1.4 Applicability to other vector-borne diseases

ArboMAP was created based on West Nile virus, however, ArboMAP can be adapted to work with any data sets that meet the following conditions:

- There are multiple years of infection data from years in which the pathogen can be considered endemic. Introductory years in which the pathogen is probably rapidly exploiting naive populations are likely not representative of transmission dynamics in subsequent years, and should be removed from the dataset.
- The pathogen has distinct transmission and quiescent seasons, with an initial growth phase at the beginning of the transmission season when the pathogen is beginning to spread after a period with minimal to no transmission. In the current version, ArboMAP assumes that the transmission season occurs during a single calendar year and ceases during the boreal winter. In settings where a transmission season crosses the December-January boundary, the code will require modification.
- There is reason to believe that incidence responds to measurable environmental indices, typically including temperature and some measure of moisture in the environment (e.g. precipitation or humidity)
- Cases of disease are assigned to districts (e.g., counties) and cases are not too rare - the best situation is where every modeled district has had at least one case over the period of study.
- Some measure of pathogen in the environment is available; e.g. here we use the rate at which pools of mosquitoes test positive for the virus.

ArboMAP is designed to facilitate statewide forecasting of West Nile virus risk at the county level, and this guide is primarily focused on this implementation.

2 Set-up and how-to guides

ArboMAP system preparation and maintenance has three phases (Figure 2).

1) An initial install to set up the software, location-specific data, and basic parameters. This will only need to be done once.

2) An annual update prior to the start of the new WNV transmission season. Previous year human data and off-season weather data will need to be updated, as well as applying any relevant software updates.

3) Immediately before running a new weekly report, only the latest mosquito pool data (if available) and weather data need to be updated.

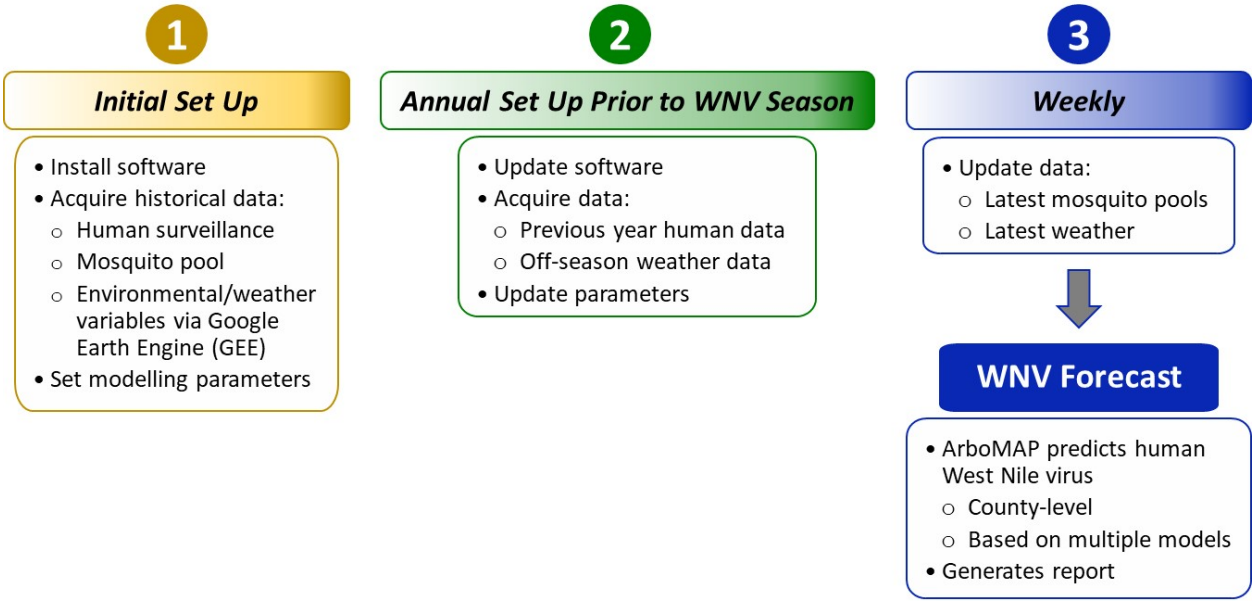

Figure 2: Three phases of ArboMAP set-up and maintence for running weekly reports: 1) Initial install, 2) Pre-season update, 3) Weekly report preparation.

2.1 Prepackaged tutorial data

ArboMAP comes packaged with example 1) human data, 2) mosquito (including optional spatial strata), and 3) weather data.

The weather data are real, but the human and mosquito data were generated randomly according to model estimates from EcoGRAPH's WNV study in South Dakota. **While the data appear similar to real human and mosquito data, they are *synthetic* and *should not be used* for any actual scientific study**. They are included merely to make sure that all software is installed and settings are correct; the user should try to run the system first with these synthetic data before trying new data.

2.2 ArboMAP data requirements & parameters

ArboMAP uses three main sources of data: human case, mosquito pool, and weather data. Historical data will need to be acquired for each. Each data format is described in the following sections, along with a

detailed explanation of the report parameters. Each data source has its own folder within the ArboMAP project (Figure 3).

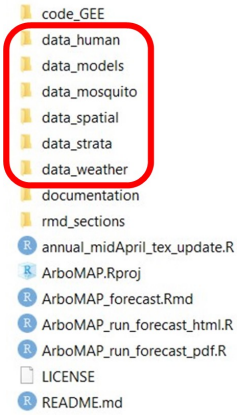

Figure 3: The data folders within the ArboMAP project, highlighted inside of the red box.

To link these datasets together and with the spatial data (county maps), there must be an identifier that is common to these datasets. ArboMAP can use ***either*** the name of the county ***or*** the FIPS code for the county. Currently these must be consistent across the first three dataset (i.e. all use name or all use FIPS codes, it does not support a mix of ID types).

FIPS is the default ID type, if it is available in human, mosquito, and weather data. If weather data has been downloaded using the ArboMAP GEE script version 2.2 or later, then FIPS codes will be included. The field names allowed are: `fips`, `FIPS`, `fips_code`, or `FIPS_CODE`.

Otherwise, it will use county names to link the datasets. There is an internal function to attempt to standardize capitalization and punctuation, however it may not succeed in all situations. The field names allowed are: `county`, `district`, `parish`, or `Parish`.

The spatial data are downloaded TIGER census shapefiles, which have a standard format and ID matching happens internally within ArboMAP.

Important information regarding dates

Correct reading of dates is notoriously difficult in the data processing field. Starting in version 4.0, ArbmoMAP will first attempt to read dates as the version 3 standard of "MM/DD/YYYY" (pattern: "%m/%d/%Y"). Should that fail (results in NA as not processable), it will try to apply non-ambiguous date formats (e.g. YYYY-MM-DD).

The date field must be present in both the human and mosquito pool data. The column must be named `date`. For human data this represents the onset date of symptoms (or whatever is used as the date of the case), and mosquito data this should be the date the mosquito pool was collected. For the weather data, there must be a `year` field and a `doy` (day of year, 1 - 366) field, which are included in the ArboMAP GEE output.

ArboMAP internally converts everything to CDC/MMWR epiweeks as the standard for the weekly report.

2.2.1 Human case data

Human case data is the outcome variable that ArboMAP is using to model and predict positive county-weeks. The availability of these data drives WNV modeling - the system can only train the model of years where there are human data and at minimum, also weather data.

Do not include any human case data from the year you are modeling. WNV cases are generally reported with large time lags of weeks to months. This would greatly underestimate the actual case counts in near-real time views, and would greatly negatively affect the modeling and forecasting accuracy.

For example, if predictions are being made for 2018, the model should be based on all human case data up to December 2017. If any human case data for 2018 is included - for example, if it is know that there were a handful of cases already in mid-July 2018 and update the human case data file during the season - then the model will assume that human case data for 2018 are complete. This is almost certainly incorrect, since some cases are only reported weeks or months after diagnosis, and human case counts during the season will therefore underestimate the actual disease burden.

ArboMAP expects a single comma-separated value (CSV) file (Figure 4), in the `data_human` folder, with these two fields:

- County identification field: This will either be the county FIPS code or the county name. See the paragraphs at the beginning of the ArboMAP data requirements & parameters section. The FIPS code field names can be `fips`, `FIPS`, `fips_code`, or `FIPS_CODE`, and the name field names can be `county`, `district`, `parish`, or `Parish`.
- `date`: the onset date of the symptoms of the case. This can be in the ArboMAP version 3 standard format of "MM/DD/YYYY" (pattern: "%m/%d/%Y"), or in non-ambiguous formats such as YYYY-MM-DD.

Other fields can be present (though please avoid having partial data in any additional ID field, as this may lead to unexpected results).

	A	B
1	district	date
2	meade	6/27/2004
3	brown	7/4/2004
4	kingsbury	7/4/2004
5	gregory	7/4/2004
6	greg0ry	7/4/2004

Figure 4: Demo human case data, also showing an intentional typo in a county name ('0' in greg0ry) which will be flagged and reported in the forecast report under the input data summary section.

2.2.2 Mosquito pool data

ArboMAP uses data on mosquito pool WNV test results to model mosquito infection rates as predictors in the model. This data is often acquired through mosquito surveillance programs implemented at the state, county, and or municipal level. The three pieces of information that are required are: the county where each mosquito pool was collected, the date that it was collected, and whether it tested positive or negative for WNV. Negative pool information (i.e. pools that test negative) are crucial data and *must* be included.

ArboMAP expects a single comma-separated value (CSV) file (Figure 5), in the `data_mosquito` folder, with these three fields:

- County identification field: This will either be the county FIPS code or the county name. See the paragraphs at the beginning of the ArboMAP data requirements & parameters section. The FIPS code field names can be `fips`, `FIPS`, `fips_code`, or `FIPS_CODE`, and the name field names can be `county`, `district`, `parish`, or `Parish`.

- `col_date`: the date of collection of the mosquito pool. This can be in the ArboMAP version 3 standard format of "MM/DD/YYYY" (pattern: "%m/%d/%Y"), or in non-ambiguous formats such as YYYY-MM-DD.
- `wnv_result`: A value of 1 for a positive test and 0 for a negative test.

Other fields can be present (though please avoid having partial data in any additional ID field, as this may lead to unexpected results). The demo file contains fields for "pool size" and "species", as these are often included in mosquito data results, but are not currently used in ArboMAP.

	A	B	C
1	district	col_date	wnv_result
2	codington	6/3/2004	0
3	codington	6/4/2004	0
4	codington	6/5/2004	0
5	codington	6/7/2004	0

Figure 5: Demo mosquito pool data showing example mandatory fields: `district` for ID here, `col_date`, `wnv_result`.

2.2.2.1 Mosquito stratification Optional: Only if a stratified mosquito model is used, then a spatial stratification file must be provided. An example stratification file for South Dakota is provided with the demonstration dataset.

If used, ArboMAP expects a single comma-separated value (CSV) file, in the `data_strata` folder, with these two fields:

- County identification field: This will either be the county FIPS code or the county name. See the paragraphs at the beginning of the ArboMAP data requirements & parameters section. The FIPS code field names can be `fips`, `FIPS`, `fips_code`, or `FIPS_CODE`, and the name field names can be `county`, `district`, `parish`, or `Parish`.
- `strata`: a unique code for each stratum, may be numerical or string (text).

This file is not needed or used for non-stratified mosquito models.

2.2.3 Weather data

Weather data is necessary to model the relationships between disease and the environment. To forecast WNV in the United States, ArboMAP uses the gridMET dataset (https://www.climatologylab.org/gridmet.html via https://developers.google.com/earth-engine/datasets/catalog/IDAHO_EPSCOR_GRIDMET), which provides daily gridded data on meteorological variables generated through fusion of the NLDAS and PRISM datasets. The ArboMAP system includes a Google Earth Engine (GEE) tools (script and an app) as options to download these data. The GEE tool takes a state and a time range and automatically processes the meteorological data, providing daily, county-level summaries in comma-separated value (CSV) format that can be imported into ArboMAP.

For details on how to run the GEE script and app, please see the How-to: Google Earth Engine (GEE) to gather weather data section.

As weather data involves large amounts of data that are updated frequently, instead of a single file, ArboMAP is built to collate multiple files into one dataset from the `data_weather` folder (Figure 6). Date ranges may

overlap, as the code will automatically take the last updated value (most recent file) for any particular day. As gridMET data are updated from the early preliminary data that are initially released, it is a good idea to include 'extra' recent history in the weekly updates. No day should be missing, however. ArboMAP will attempt to fill in the best that it can, but it will report any missing dates in the forecast report, and the GEE script should be run to gather data for these dates.

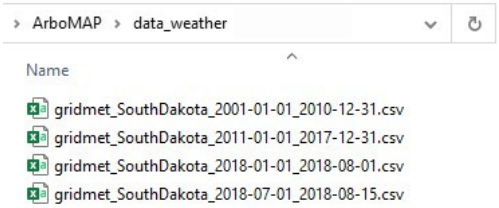

Figure 6: Multiple files are allowed in the `data_weather` folder, with overlapping dates. ArboMAP will collate all files using the most recent data for any particular day.

While GEE is very useful for a number of reasons, ArboMAP can use environmental data from any source as long as the csv files follow the format below (Figure 7).

- County identification field: This will either be the county FIPS code or the county name. See the paragraphs at the beginning of the ArboMAP data requirements & parameters section. The FIPS code field names can be `fips`, `FIPS`, `fips_code`, or `FIPS_CODE`, and the name field names can be `county`, `district`, `parish`, or `Parish`.
- `doy`: indicates the day of the year (1-366) of the observation.
- `year`: the year of the observation.
- The remaining columns are measurements of the observed environmental conditions. If the measurements are not daily, they will need to be resampled to that temporal resolution. The column name MUST match the two predictor names given in the parameters (see the Parameters section for more details). Additional columns are allowed and are not used.

	A	B	C	D	E	F	G	H	I	J	K
1	district	fips	doy	year	tminc	tmeanc	tmaxc	pr	rmean	vpd	vs
2	Codington	46029	182	2018	15.764	20.8814	25.999	1.483	61.579	1.04	4.91
3	Brown	46013	182	2018	13.979	21.6383	29.298	0	53.856	1.513	5.022
4	Edmunds	46045	182	2018	14.785	21.847	28.909	0.004	51.731	1.519	5.263
5	Sully	46119	182	2018	14.69	21.802	28.914	0.127	56.759	1.422	4.813

Figure 7: GridMET weather data downloaded from ArboMAP Google Earth Engine script.

2.2.4 Models

Models to be used in an ArboMAP run are specified in a text file - in the demo, `models.txt` in the `data_models` folder of the RStudio project. Each line in this file contains two comma-separated strings, with **no** space between the fields. The first contains a short name for the model and the second contains the specification of the R model formula.

For more details on the models in the demo, see the ArboMAP modeling section.

2.2.5 Spatial

ArboMAP uses spatial data to link the data and forecasts to their county to display in the maps. The demo comes pre-packaged with the census spatial data for South Dakota, however ArboMAP has the ability to create this for any state (given an internet connection).

In the parameters, for the spatial data file location (`file_spatial_sf`), enter the command `create` to download the census spatial data once and create a file in `data_spatial` folder. Use the command `always_download` to download census spatial data each time.

After creation (if `create` is used), the file will exist in the `data_spatial` folder and will be named like `sd_counties.RDS`, where "sd" is replaced with the state code.

Note for advanced users: This is being downloaded via the tigris package using the following command: `tigris::counties(state = params$state_code, cb = TRUE)`, where the `state_code` is one of the parameters set by the user (see the Parameters section).

2.2.6 Parameters

ArboMAP uses input parameters to control the specific configuration for a forecast report. The defaults are set within the header of the `ArboMAP_forecast.Rmd` file, however, the user can change them for every run on the fly. When running the forecast from the run scripts (see the How-to: Weekly reports section), a browser will pop up an interactive interface to change parameters (Figure 8). Not all parameters will need to be changed when running weekly reports, usually only the forecast date will be changed week to week.

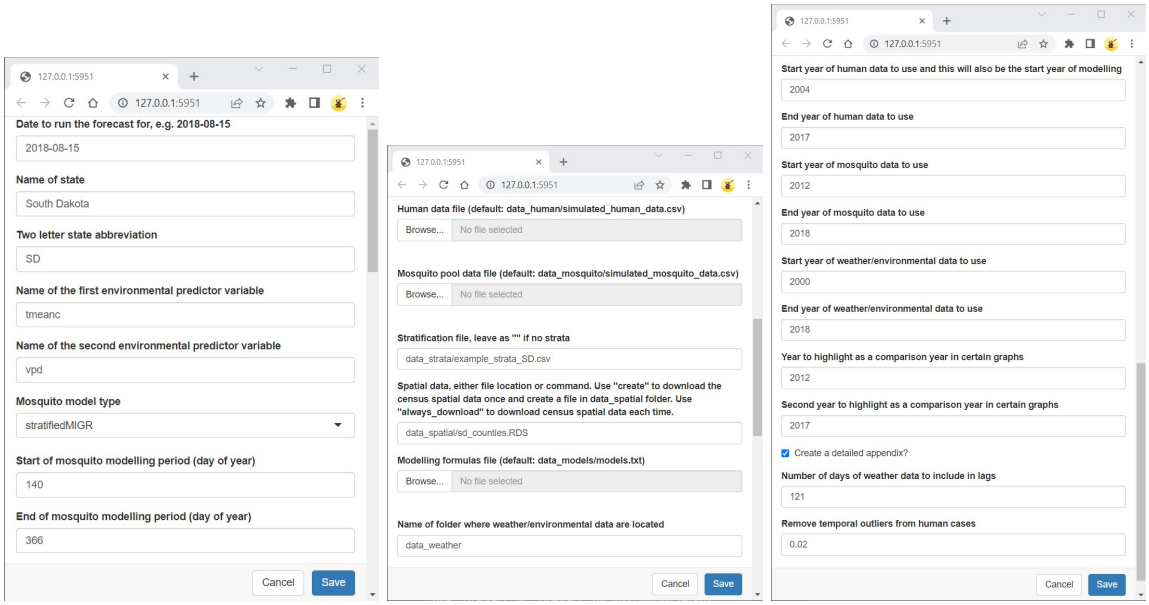

Figure 8: Browser window showing the parameter interface from run scripts. Only one window is open, three images shown here cover all parameters (scrolling down).

The parameters, listed in order of appearance, with both code names and long descriptors:

- `state_name`: The name of the state being modeled. This will be included in the title of the report generated by ArboMAP.
- `state_code`: The two-letter state abbreviation. This will be used in the creation of spatial data, if necessary.

- `forecast_date`: The requested date of the forecast. ArboMAP is based on weekly forecasts, using CDC/MMWR epiweeks, so the system will convert the requested forecast date into the epiweek it falls into. The date itself will appear in the title of the report. This will be updated each time the report is run.
- `predictor_var1` and `predictor_var2`: The names of the two meteorological variables chosen as predictors. These should match the name of the appropriate columns in the weather data. In the demo data, these are `tmeanc` for mean temperature in Celsius and `vpd` for vapor pressure deficit.
- `mosquito_model` is a code for the type of mosquito infection rate model to be used:
  - `simpleratio`: Percent positive mosquito pools
  - `AUC`: Area under the mosquito infection growth curve
  - `MIGR`: Mosquito infection growth rate
  - `MII`: Mosquito infection intercept (constant term from the mosquito infection growth rate equation)
  - `stratifiedMIGR`: Spatially stratified version of MIGR
  - `stratifiedMII`: Spatially stratified version of MII.
- `mosquito_doy_start`: Include mosquito data per year starting at this day of year. The mosquito infection rate modeling is very sensitive to early mosquito pool results, which is why a cut-off is used. Sensitivity analyses indicate that a start day of year of 140 is a reasonable cut-off for a high modeling accuracy.
- `mosquito_doy_end`: The ending day of year (per year) to include mosquito data for modeling.
- `file_human`: The path and file that contains all human case data.
- `file_mosquito`: The path and file that contains all mosquito pool WNV testing data.
- `file_strata`: Optional path and file for dividing the counties into different strata for mosquito modeling and only used with stratified mosquito models.
- `file_county_sf`: Spatial data, either a file location or command. Use `create` to download the census spatial data once and create a file in the `data_spatial` folder that can be used in all subsequent runs. Use `always_download` to download census spatial data each time.
- `file_models`: Modelling formulas file, comma separated (with no space between fields) text (txt) file.
- `folder_weather`: Folder where weather/environmental data files are located. Subfolders will *not* be included.
- `year_human_start`: Start year of human data to use. As modelling cannot happen without outcome data (human cases), this will also be the start year of modelling.
- `year_human_end`: End year of human data to use. Unless you are running some kind of special tests or reports, this should be the year ***before*** the current forecast year. See the Human case data section for a discussion on this.
- `year_mosquito_start`: Start year of mosquito data to use.
- `year_mosquito_end`: End year of mosquito data to use. This is normally the same year as the current forecast year, or the last year of available mosquito information.
- `year_weather_start`: Start year of weather/environmental data to use. Will want to have at least lag length number of days before the start of modelling (`year_human_start`), so generally at least the year before that. However, more can be included and the additional data will be in the calculation of the historical statistics (historical medians, etc.).
- `year_weather_end`: End year of weather/environmental data to use. This is normally the same year as the current forecast year.
- `year_compare_vis1` and `year_compare_vis2`: Two years to highlight as a comparison year in certain graphs.
- `create_appendix`: Check the box to create a detailed appendix that will create output PER model, rather than the model averages. (Note for advanced users: this is TRUE or FALSE in the Rmd file.)
- `lag_length`: Number of days of weather data to include in lags (for the delayed effects of weather data on disease transmission risk). ArboMAP will use the previous number of days, e.g. previous 121 days, of environmental data of each date to include in the smoothed term in the models.
- `case_trim_alpha`: Remove temporal outliers from human cases. This removes the earliest & latest `case_trim_alpha` % of human cases over all years. This prevents the outliers, which may be cases

with incorrect onset dates, from causing numerical difficulties in the modelling.

2.3 How-to: Google Earth Engine (GEE) to gather weather data

ArboMAP system includes Google Earth Engine (GEE) tools to download weather data. ArboMAP uses the gridMET dataset (https://www.climatologylab.org/gridmet.html via https://developers.google.com/earth-engine/datasets/catalog/IDAHO_EPSCOR_GRIDMET), which provides daily gridded data on meteorological variables generated through fusion of the NLDAS and PRISM datasets. The GEE tool (script or app) takes a state and a time range and automatically processes the meteorological data, providing daily, county-level summaries in a comma-separated value format that can be imported into ArboMAP.

There are two main ways to interface with the ArboMAP GEE code:

1) as a javascript script through the GEE Code Editor website, or
2) as an app through a website (which is running the same code in the background).

Either one allows the user to select a state and a date range to generate and download the summaries.

The Code Editor method allows for longer date ranges (multiple years), requires a free GEE account, and the data is downloaded to the user's Google Drive when it is ready. It is the recommended way for the initial data set up for a new project.

The GEE app is designed for users without a GEE account, can handle requests up to several months, and downloads directly from the browser to the user's computer. It can be used for the annual off-season or weekly updates.

2.3.1 Code Editor version

There will be a `arbomap_gridmet_gee_v[version#].js` file in the `code_GEE` folder of the ArboMAP project. This will be the latest released version of the ArboMAP GEE code (with appropriate version number in the file name).

2.3.1.1 Set-up

1) Navigate to https://code.earthengine.google.com.
2) In the "New Script" center section of the page, copy and paste the text of the script.
3) Click on "Save", and name it the same as the .js file.

Alternatively, from the app (https://dawneko.users.earthengine.app/view/arbomap-gridmet), follow the link "Access to Code Editor version (must have Google Earth Engine account)" near the bottom of the left-hand pane. This will bring the user to a snapshot of the code, which can be copied into a new file owned by the user, or used as is.

2.3.1.2 To run

1) Click on the "Run" button in the top-middle pane (Step 1 in Figure 9). This will bring up the user interface in a panel on the left-side.
2) Under the "Downloader" section, select the state from the drop down list (Step 2.1 in Figure 9).
3) Change the download start and end dates as necessary (Step 2.2 in Figure 9). As gridMET data are updated from the early preliminary data that are initially released, it is a good idea to include 'extra' recent history in the weekly updates. ArboMAP will collate the data and use the latest updated values automatically. The default end date is the date of latest available data, and the default start date is a month prior.

4) In the "Task" pane, click on the blue button labeled "RUN" to kick off the processing (Step 3 in Figure 9). Note that while a popup download link appears on top of the map, it should not be used for the long time ranges (it will timeout), and should be ignored when doing so via the Code Editor.

5) The file will be saved to the user's Google drive when finished. Move the file into the `data_weather` folder of the ArboMAP project.

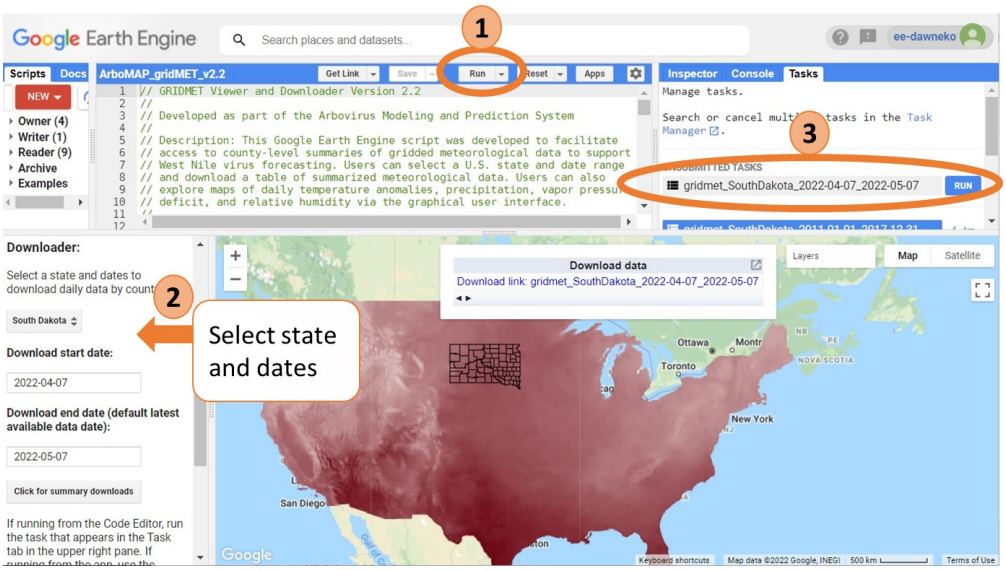

Figure 9: Code Editor interface for the ArboMAP Google Earth Engine script to download gridMET weather data. The primary steps are labeled with key items marked.

As weather data involves large amounts of data that are updated frequently, instead of a single file, ArboMAP is built to collate multiple files into one dataset from the `data_weather` folder (Figure 6). Date ranges may overlap, as the code will automatically take the last updated value (most recent file) for any particular day. No day should be missing, however. ArboMAP will attempt to fill in the best that it can: explicit missing will be filled in with modeled values during anomaly calculations and implicit missing with historical medians. However, it will report any missing dates in the forecast report, and the GEE script should be run to gather data for these dates.

2.3.2 App

1) Navigate to https://dawneko.users.earthengine.app/view/arbomap-gridmet (Figure 10).

2) Under the "Downloader" section in the left-hand pane, select the state from the drop down list (Step 1 in Figure 10).

3) Change the download start and end dates as necessary (Step 2 in Figure 10). As gridMET data are updated from the early preliminary data that are initially released, it is a good idea to include 'extra' recent history in the weekly updates. ArboMAP will collate the data and use the latest updated values automatically. The default end date is the date of latest available data, and the default start date is a month prior.

4) In the small window that pops-up on the map, click the download link, which will download the file directly from the browser to your computer (Step 3 in Figure 10). Save or move the file into the `data_weather` folder.

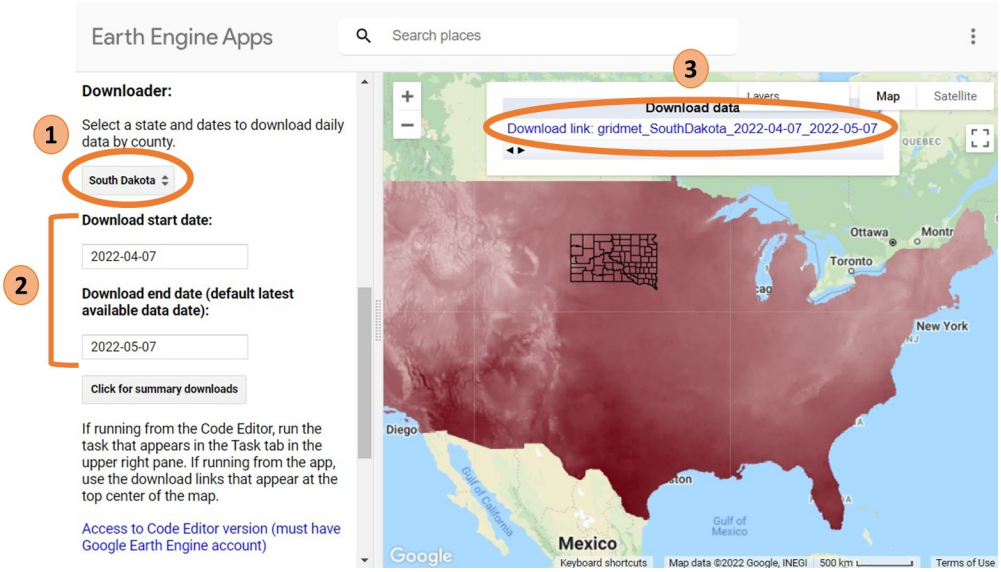

Figure 10: ArboMAP Google Earth Engine app to download gridMET weather data. The primary steps are labeled with key items marked.

As weather data involves large amounts of data that are updated frequently, instead of a single file, ArboMAP is built to collate multiple files into one dataset from the `data_weather` folder (Figure 6). Date ranges may overlap, as the code will automatically take the last updated value (most recent file) for any particular day. No day should be missing, however. ArboMAP will attempt to fill in the best that it can: explicit missing will be filled in with modeled values during anomaly calculations and implicit missing with historical medians. However, it will report any missing dates in the forecast report, and the GEE script should be run to gather data for these dates.

2.4 Set-up: Initial install

2.4.1 R

R (https://www.r-project.org) is a statistical programming language that runs all of our analyses and produces reports and documentation, including this document. It is free and has a wide variety of packages built by users all around the world to do to do many different statistical analyses with graphing and display options.

It is easiest to download R from the Comprehensive R Archive Network (CRAN): https://cloud.r-project.org. For Windows user, please click on the link for "Download R for Windows" (Figure 11). Choose "base," then "Download R for Windows" (Figure 12). Run this file and install R on your system. Use the default settings for the installer.

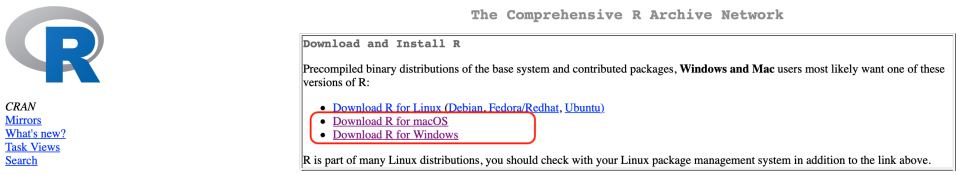

Figure 11: Main download link page on CRAN for R

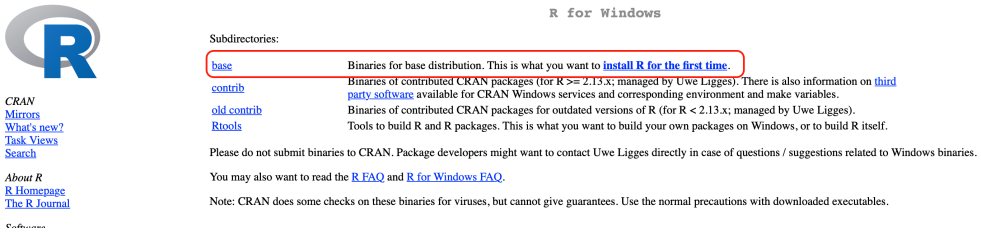

Figure 12: For Windows, pick base to install R for the first time

For Mac users, please click on the link "Download R for macOS." Choose latest released R. Use the default settings for the installer.

2.4.2 RStudio

RStudio (https://www.rstudio.com) is a user-friendly GUI for the statistical programming language R that greatly simplifies a number of tasks for the programmer. Navigate to the site, click on "Download", and choose the "RStudio Desktop" option (Open Source License; the free version). Run the appropriate installer for your system.

2.4.3 R packages

This is an OPTIONAL step. This will happen automatically the very first time you run a forecast, but if you wanted to reduce the amount of time that would take, this is a good preparatory step.

An R package is a collection of functions, data, and documentation that extends the capabilities of base R. Many packages are used in our code for forecasting, and will need to be installed.

In the RStudio console, run the command below and that will install a good number of the packages that are used from the `tidyverse` set of packages, plus others used for data processing and report generation.

```
install.packages(c("tidyverse", "tidyselect", "glue", "zoo", "mgcv", "splines",
                   "parallel", "pROC", "tigris", "sf", "gridExtra", "viridis",
                   "ggrepel", "ggpubr", "knitr", "rmarkdown", "shiny"))
```

Note to advanced users: If a pdf report is being generated, and the computer does not have a LaTeX installation, it will automatically download and install TinyTeX. If R can detect a LaTeX installation, it will skip this step, and the user should verify that their Sweave options in RStudio are set appropriately.

2.4.4 ArboMAP

Please download the latest ArboMAP project ('Source code') from the Github repository of the EcoGRAPH Research Group (https://github.com/EcoGRAPH/ArboMAP/releases) and extract or unzip the contents.

ArboMAP comes packaged with 1) example human data, 2) example mosquito data, including optional spatial strata, and 3) weather data.

The weather data are real, but the human and mosquito data were generated randomly according to model estimates from an EcoGRAPH WNV study in South Dakota. **While the data appear similar to real human and mosquito data, they are *synthetic* and *should not be used* for any actual scientific study**. They are included to make sure that all software is installed and settings are correct. The user should test run the system first with these data before changing to other data sources.

Adapting ArboMAP to a different data set and location will involve:

- Changing the human, mosquito pool, and weather data sets.
- Modifying the parameters to the appropriate settings.

See ArboMAP data requirements & parameters section above for details on data requirements & formatting, and a description of all parameters.

2.4.5 Google Earth Engine

Environmental data will be obtained from Google Earth Engine (GEE). GEE is a cloud-based platform for hosting satellite imagery. GEE also provides tools to process these remote sensing images and other geospatial datasets. Instead of downloading the raw satellite files and processing them on the user's own computer, which requires significant internet bandwidth and processing power, these steps are done in the cloud. At the end, only the summarized output will be downloaded.

There are two options to gather environmental data in GEE for ArboMAP. The app, accessed via a website, does not require a Google Earth Engine account, but is limited to short periods (months, not years) of data at a time (https://dawneko.users.earthengine.app/view/arbomap-gridmet). The script, accessed through the GEE Code Editor website, allows for longer time periods, and does require a GEE account, which can be set-up using the following instructions:

1) Request a GEE account: sign up at https://earthengine.google.com. If the user does not already have a Google Account, it will prompt to make one. The Google account will also contain a Google Drive account, which is where the GEE data will be downloaded to.

2) Copy the GEE code into a new script and save in the user's account. The code can be found in the `arbomap_gridmet_gee_v[version#].js` file in the `code_GEE` subfolder of the ArboMAP project download.

The GEE code may alternatively be found by following the "Access the Code Editor version" link from the ArboMAP gridMET GEE app https://dawneko.users.earthengine.app/view/arbomap-gridmet. It can be run from this fixed link, or copied into a new script owned by the user (as in step 2a).

Details on how to use GEE are found in the How-to: Google Earth Engine (GEE) to gather weather data section.

2.5 Set-up: Annual update at start of WNV season

When preparing to start running weekly reports for the new WNV season, there are three main steps that are needed annually: 1) updating software, 2) acquiring data, and 3) updating parameters.

2.5.1 Annual software update

When pdf reports are created, ArboMAP is using tinytex in the background. TeX Live is updated every year at the beginning of April, so ArboMAP must also be updated.

1) Click on the `ArboMAP.Rproj` file to open the project in RStudio.
2) In RStudio, in the Files pane, click on `annual_midApril_tex_update.R` (Figure 13).

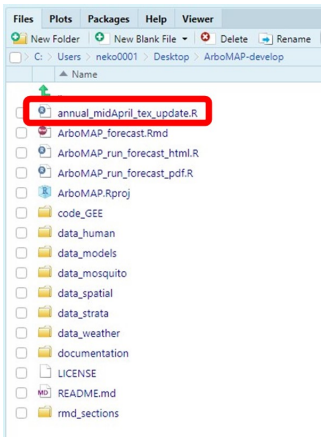

Figure 13: RStudio Files pane showing the annual TeX update script marked.

3) In the Source pane that opens up, make sure the cursor is at the start of line 1, and hit "Run" (Figure 14).

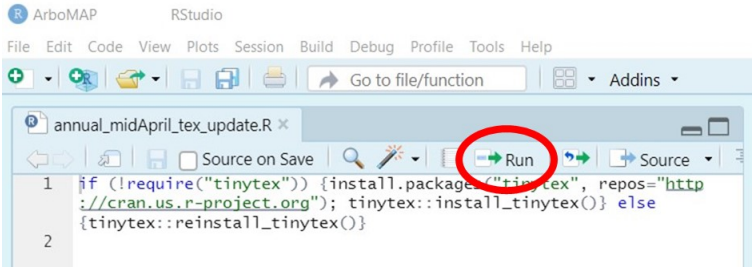

Figure 14: Annual TeX update script as seen in RStudio Source pane.

2.5.2 Annual data update

2.5.2.1 Human data ArboMAP does not (and should not) use human case data from the current forecast year, due to the large time delays in diagnosis and reporting of West Nile virus infections, which would greatly underestimate the actual case counts in near-real time views.

At the beginning of a new season, i.e. a new forecast year, the previous year's human case data will need to be added to the data file. E.g. in the spring of 2019, the human case data for 2018 should be appended to the human case csv file. (Or a new extract from the original database, whatever if appropriate.)

ArboMAP expects a single comma-separated value (CSV) file (Figure 4), in the `data_human` folder, with these required fields:

- County identification field: This will either be the county FIPS code or the county name. See the paragraphs at the beginning of the ArboMAP data requirements & parameters section. The FIPS code field names can be `fips`, `FIPS`, `fips_code`, or `FIPS_CODE`, and the name field names can be `county`, `district`, `parish`, or `Parish`.
- `date`: the onset date of the symptoms of the case. This can be in the ArboMAP version 3 standard format of "MM/DD/YYYY" (pattern: "%m/%d/%Y"), or in non-ambiguous formats such as YYYY-MM-DD.

Other fields can be present, but are not used. See the Human case data section for more details if needed.

2.5.2.2 Weather data At the start of a new WNV transmission risk season (e.g. spring), the weather data will need to be updated from since last run, e.g. the last weekly report run of the previous transmission risk season (e.g. previous fall).

Use the GEE app (or other source, if relevant) to gather weather data from the end of last forecast season (or whenever weather data was last updated) through current.

1) Navigate to https://dawneko.users.earthengine.app/view/arbomap-gridmet (Figure 15).
2) Under the "Downloader" section, select the state from the drop down list (Step 1 in Figure 15)
3) Change the download start and end dates as necessary (second part of Step 2 in Figure 15). A good start date would be a week or so before the last known previous data. E.g. if the weather data was last updated to 2022-10-03, it is reasonable to run this annual update from 2022-09-03 or so. This ensures that the ArboMAP will have the final updated gridMET data, and not the early released preliminary data.
4) In the small window that pops-up on the map, click the download link, which will download the file directly from the browser to your computer (Step 3 in Figure 15). If it times out, split the download into two time ranges, or use the GEE script method instead.

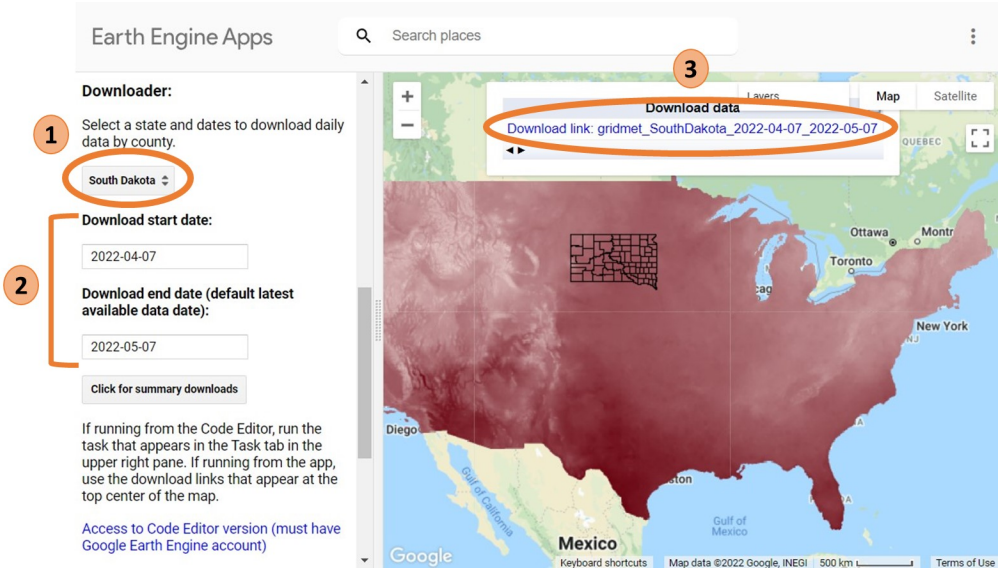

Figure 15: ArboMAP Google Earth Engine app to download gridMET weather data. The primary steps are labeled with key items marked.

For more details, see the How-to: Google Earth Engine (GEE) to gather weather data section.

2.5.3 Annual parameter updates

End year and potentially file names will need to be updated at the beginning of a new season (year).

The defaults can be updated by manually editing the header of the `ArboMAP_forecast.Rmd` file, or the correct setting or file can be picked from the interactive user interface each time.

- `year_mosquito_end`: Update the end year of mosquito data to use. This normally is the same year as the current forecast year.
- `year_weather_end`: Update the end year of weather/environmental data to use. This normally is the same year as the current forecast year.

IF the file names of the human or mosquito data has changed, those parameters will also need to be updated.

- `file_human`: The path and file that contains all human case data.
- `file_mosquito`: The path and file that contains all mosquito pool WNV testing data.

See the Parameter section for a description of all parameters.

2.6 How-to: Weekly reports

2.6.1 Update data

2.6.1.1 Mosquito data The latest mosquito pool data with positive or negative WNV test results will need to be appended to the mosquito data file. (Or a new extract from the original database, whatever if appropriate.)

ArboMAP expects a single comma-separated value (CSV) file (Figure 5), in the `data_mosquito` folder, with these fields:

- County identification field: This will either be the county FIPS code or the county name. See the paragraphs at the beginning of the ArboMAP data requirements & parameters section. The FIPS code field names can be `fips`, `FIPS`, `fips_code`, or `FIPS_CODE`, and the name field names can be `county`, `district`, `parish`, or `Parish`.
- `col_date`: the date of collection of the mosquito pool. This can be in the ArboMAP version 3 standard format of "MM/DD/YYYY" (pattern: "%m/%d/%Y"), or in non-ambiguous formats such as YYYY-MM-DD.
- `wnv_result`: A value of 1 for a positive test and 0 for a negative test.

See the Mosquito data section for more details if needed.

2.6.1.2 Weather data Use the GEE app (or other source, if relevant) to gather the most recent weather data.

1) Navigate to https://dawneko.users.earthengine.app/view/arbomap-gridmet (Figure 16).
2) Under the "Downloader" section, select the state from the drop down list (Step 1 in Figure 16).
3) Change the download start and end dates if necessary (second part of Step 2 in Figure 16). The default end date is the last available data and does not need to be changed. The default start date is one month prior, and this is a reasonable timeframe to make sure to get the latest gridMET updated data (which initially released preliminary data and later updates to final data).
4) In the small window that pops-up on the map, click the download link, which will download the file directly from the browser to your computer (Step 3 in Figure 16).

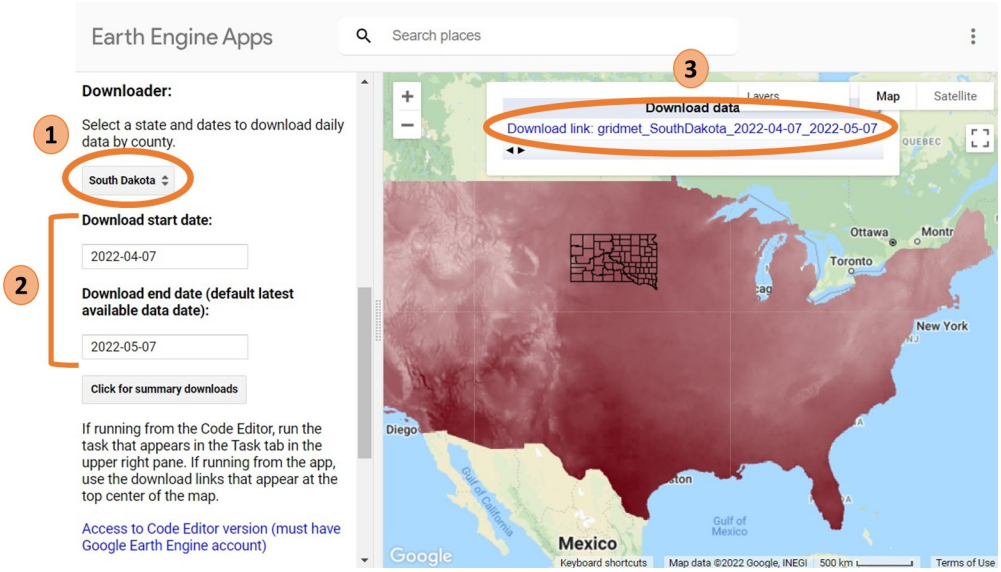

Figure 16: ArboMAP Google Earth Engine app to download gridMET weather data. The primary steps are labeled with key items marked.

For more details, see the full GEE instructions.

2.6.2 Run forecast

There are a few options for how to kick off a forecast report. The easiest ways are to use the appropriate run script.

1) In the ArboMAP folder, click on `ArboMAP.Rproj` file. This will open the ArboMAP project in RStudio.

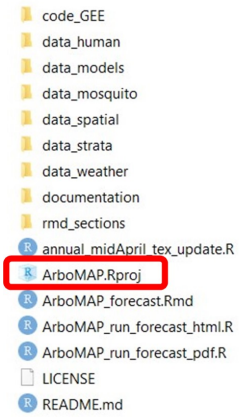

Figure 17: To open the ArboMAP project in RStudio, click on the ArboMAP.Rproj file.

2) In the File pane of RStudio, click on the appropriate run script.

- `ArboMAP_run_forecast_pdf.R`: a pdf file of the forecast report will be created

- `ArboMAP_run_forecast_html.R`: an html file of the forecast report will be created

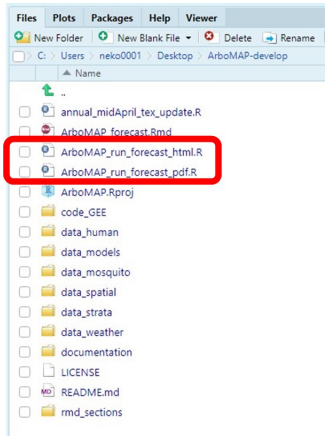

Figure 18: Pre-built run scripts to create either html or pdf forecast reports. Clicking and running these will open an interactive user interface to adjust parameters.

3) In the Source pane that opens, make sure the cursor is on line 1 and hit the 'Run" button. The pdf run script is shown in Figure 19, and the html run script will look very similar.

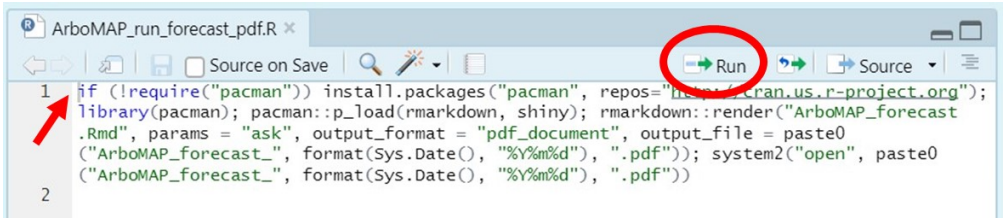

Figure 19: Source pane for the pdf run script, indicating that the cursor should be on line 1 and the location of the Run button.

4) This will pop open a browser window for a interactive interface for the parameters.

- Change the forecast date to today's date (or date of interest). ArboMAP is based on weekly forecasts, using CDC/MMWR epiweeks, so the system will convert the requested forecast date into the epiweek it falls into. The date itself will appear in the title of the report.
- If there are other parameters that need to be changed (e.g. from annual updates that were not changed in the default settings), then remember to change those as well here.

5) Click on 'Save' to kick off the report in RStudio. After it finishes running, it will pop open the file to view. It will automatically add today's date to the filename, rename the file if desired.

Note: The very first time a report is run on a computer, it will spend some time downloading and installing necessary R and TeX packages. See the R packages section for a command to run prior to a report to pre-install most of the necessary R packages.

3 Report interpretation

ArboMAP produces a weekly, county-level forecast of human West Nile virus (WNV) cases using environmental data combined with entomological data, in either pdf or html format depending on user choice. The report always includes a main section, which reports the forecast results as an ***average*** of all models. An optional appendix (set in user parameters) will display results ***per model***, as well as some additional analyses and details.

3.1 Forecast results

3.1.1 Absolute risk

The absolute risk maps display the risk of a county having at least one positive case during the forecast epidemiological week (epiweek). The main report will show the average risk from all models (Figure 20) and the appendix will show a map per model.

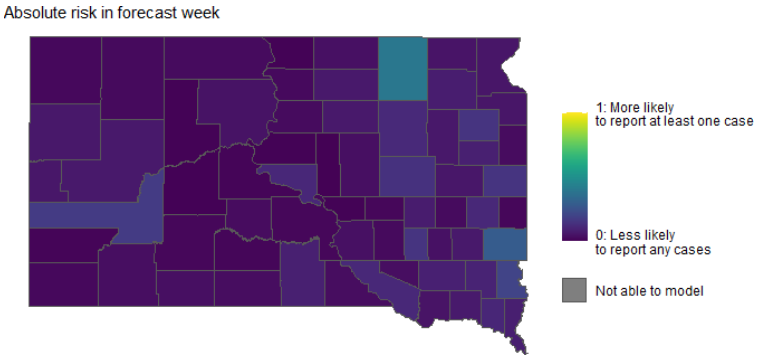

Figure 20: Demonstration absolute risk map for South Dakota.

The absolute risk map can be used in conjunction with the relative risk map. The absolute risk map shows the risk of a county reporting at least one WNV positive human case during this week, and the relative risk map shows if this risk is elevated (or not) as compared to previous years.

3.1.2 Relative risk

The relative risk maps display the risk of a county having at least one positive case *compared to* the county's risk in previous years during the same epiweek. The main report will show the average risk from all models (Figure 21) and the appendix will show a map per model.

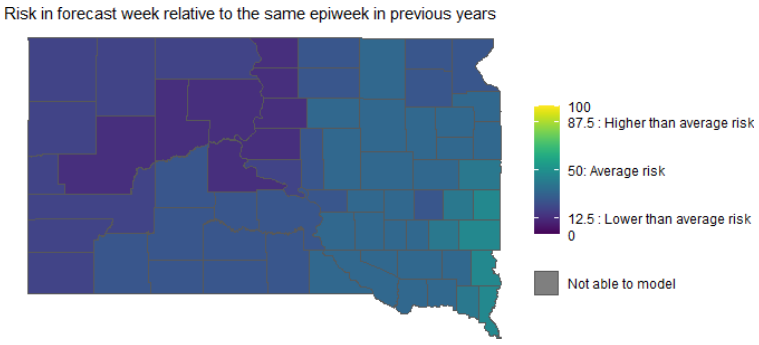

Figure 21: Demonstration relative risk map for South Dakota.

The relative risk map can be used in conjunction with the absolute risk map. The absolute risk map shows the risk of a county reporting at least one WNV positive human case during this week, and the relative risk map shows if this risk is elevated (or not) as compared to previous years.

If there are counties with greater than average risk (defined as >87.5%), then it will be listed in a table that appears after the map.

3.1.3 Current and multi-year forecasting charts

In two different sections, the predicted epicurves are shown in timeseries graphs: one for just the current forecast year, and another for all modeled years.

In the main report, the graphs show the average of all models (dark red line), with the range of all models in the shaded ribbon (Figures 22 and 24). In the appendix, the graphs will show each model as a separate series (various colors; Figures 23 and 25)). In all, the forecasts are shown as a dotted line and the predicted values from before the current forecast week ('backcast') are shown as a solid line.

The current forecast year graphs also show the observed proportion of counties positive averaged from all known years (Figures 22 and 23). This is excluding human cases that occurred very early or very late in the season (temporal outliers), based on the percentage cut-off in the parameters (`case_trim_alpha`). This plotted curve allows a comparison between the timing and height of the predicted peak of cases as compared to averaged historical years.

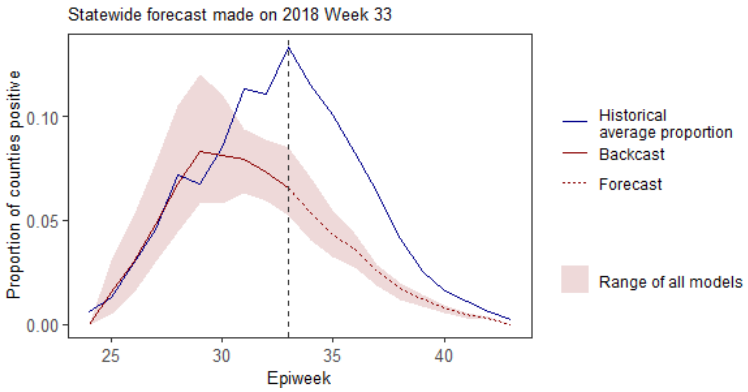

Figure 22: Demonstration predicted epicurve of the forecast year: the average of all models (red line) with ribbon for the range of the models. The historical observed proportion of counties positive, averaged over all known years, is also shown (dark blue line).

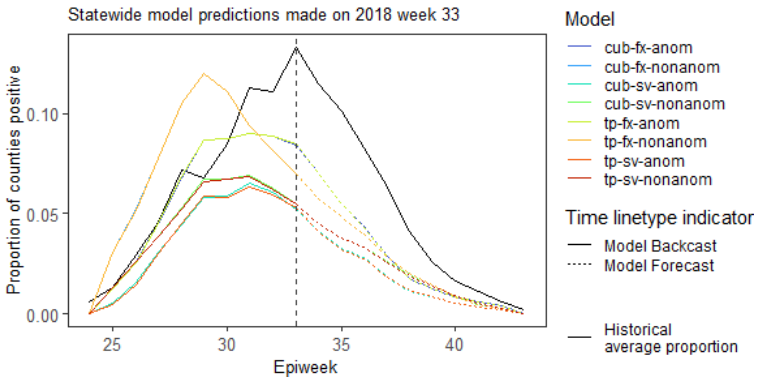

Figure 23: Demonstration predicted epicurves of the forecast year: each model is shown as a separate series in the appendix (various colors). The historical observed proportion of counties positive, averaged over all known years, is also shown (black line).

The multi-year graphs also show the historical observed values (Figures 24 and 25).

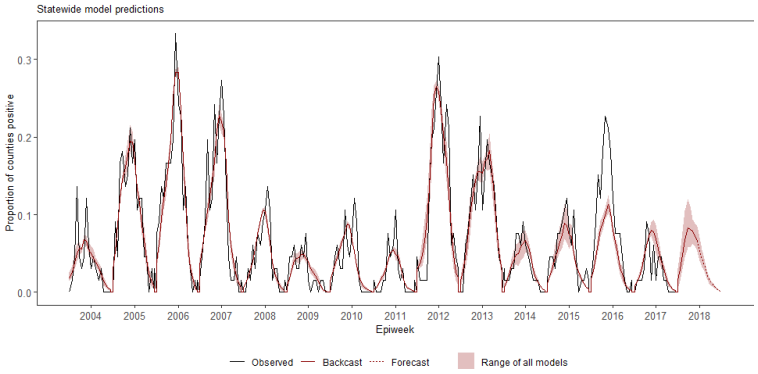

Figure 24: Demonstration predicted epicurves for the entire modeled period: the average of all models (red line) with ribbon for the range of the models. The historical observed values are also shown (black line).

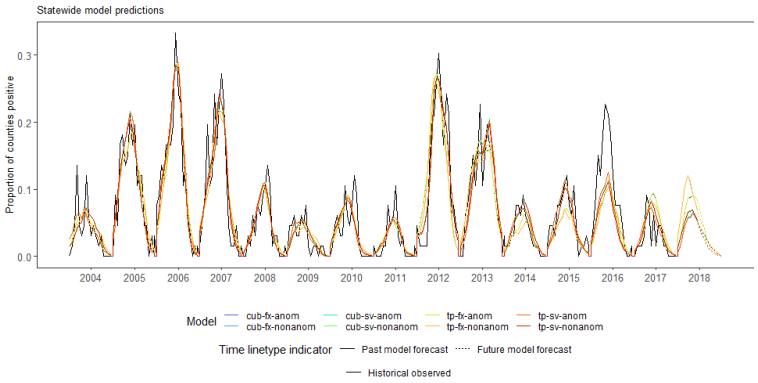

Figure 25: Demonstration predicted epicurves for the entire modeled period: each model is shown as a separate series in the appendix (various colors). The historical observed values are also shown (black line).

3.1.4 Case estimation

ArboMAP models are based on 'positive county-weeks', the probability that a county would have at least one human WNV case in a given week. These 'positive county-week' values were used to model and predict the total number of *cases*. Historical data (predicted postive county-weeks and observed cases) were used to fit a case estimation model, and predictions were made for the current forecast year. Predictions are made per model - the main report will show the average (Figure 26) and range of the models and the appendix will report each model out separately (Figure 27).

Table 1: Estimated number of WNV cases

Year	Predicted positive county-weeks	Average estimated cases (standard dev)	Range of estimated cases
2018	49	59 (+/-14)	43 - 76

Figure 26: Demonstration estimated human WNV cases from the main report showing the average and range of all models.

Table 6: Estimated number of WNV cases

Year	Model	Predicted positive county-weeks	Estimated cases
2018	cub-fx-anom	55.1	66
2018	cub-fx-nonanom	63.7	75
2018	cub-sv-anom	35.8	43
2018	cub-sv-nonanom	42.5	51
2018	tp-fx-anom	55.2	66
2018	tp-fx-nonanom	63.7	75
2018	tp-sv-anom	35.4	43
2018	tp-sv-nonanom	42.3	51

Figure 27: Demonstration estimated human WNV cases from the appendix showing each model results separately.

3.1.5 Model fit statistics

The report includes statistics on how well the model is fitting historical years (Figure 28). The main report will include the Area Under ROC Curve (AUC) for the average of the models. The appendix will report AUC, AIC, Temporal MAE, and Spatial MAE for each model:

- AUC : Area Under ROC Curve, values range from 0 (model is right 0% of the time) to 1 (model is right 100% of the time). Scores above 0.5 are better than a random model, with >0.7 generally considered acceptable and >0.8 as good.
- AIC : Akaike information criterion, relative fit statistic to other models.
- Temporal MAE : Mean Average Error, mean of weeks (collapsed to state).
- Spatial MAE : Mean Average Error, mean of counties (collapsed all time).

Table 2: Area Under Curve (AUC) statistics of all model fits

Model	Average AUC	Min AUC	Max AUC
Average of all models	0.85	0.84	0.86

Table 7: Fit statistcs by model

Model	AUC	AIC	MAE Temporal	MAE Spatial
cub-fx-nonanom	0.845	6663	1.688	0.741
cub-fx-anom	0.844	6676	1.649	0.732
cub-sv-nonanom	0.854	6544	1.472	0.708
cub-sv-anom	0.855	6567	1.409	0.698
tp-fx-nonanom	0.845	6663	1.689	0.741
tp-fx-anom	0.844	6675	1.648	0.732
tp-sv-nonanom	0.854	6544	1.471	0.708
tp-sv-anom	0.855	6566	1.408	0.698

Figure 28: Demonstration model fit statistics. Table on the top is from the main report showing the average and range of all models, and the table on the bottom is from the appendix showing additional statistics for each model results separately.

3.1.6 Partial effects

In the appendix only, there will be a section for partial effects plots for each model. ArboMAP allows the user to write custom model formulas, and as such the plots in this section are the partial effects of all the smooth terms for each model. The plots show the component effect of each smooth term. All components (not just smooths) added together would be the overall prediction. For a table of all demo formulas, see the ArboMap modeling section.

The number after the comma in the `s({item}, {number})` labels is the effective degrees of freedom (EDF). The EDF is a measurement of the complexity of the smooth term - a value of 1 is a straight line, higher values are more complex curves.

An easy way to check on the significance of the smooth term is if a horizontal line cannot be drawn through the 95% confidence interval (value +/- se, shown in the gray shaded ribbon in the relevant graphs).

- s(doy): a smooth over the day of the year (doy). This is usually a cyclical term used in anomalized models, which models the general trend over all years in the sample. This is relative - the lower the function at some day of the year, the less likely you are to see human WNV cases on that day of the year, all else being equal. Generally, this should range between -5 and 5 - if it goes beyond this range, it is likely that too many zero case county-weeks are being included in the model and parameter `case_trim_alpha` should be increased so that more empty weeks are removed.
- s(lag):a single distributed distributed lag function, showing the dependence of risk now (lag = 0) on environmental data some time in the past. As an example, if the estimated distributed lag is positive at lag = 60, then the environmental variable two months ago correlates with an increase in human cases today.
- te(lag, doymat): a variable that indicate how the risk today (lag = 0) depends on environmental data at some point in the past. If a seasonally-varying model was chosen, then these will depend on the day of the year and three different dependence functions will be shown. For example (not shown), precipitation in the winter (near doy = 0) will not have much of an effect on human risk, whether two days (lag = 2) or three months (lag = 90) ago. If these lines are essentially all the same, a better fit may be achieved with fixed (fx) rather than seasonally-varying (sv) distributed lags.

All models with smooths will have 1-D graphs (e.g. Figure 29). Seasonally-varying models will also have 2-D graphs (components with `doymat` in standard models), however a subset of y-values have been pulled out to plot as lines (e.g. Figure 30).

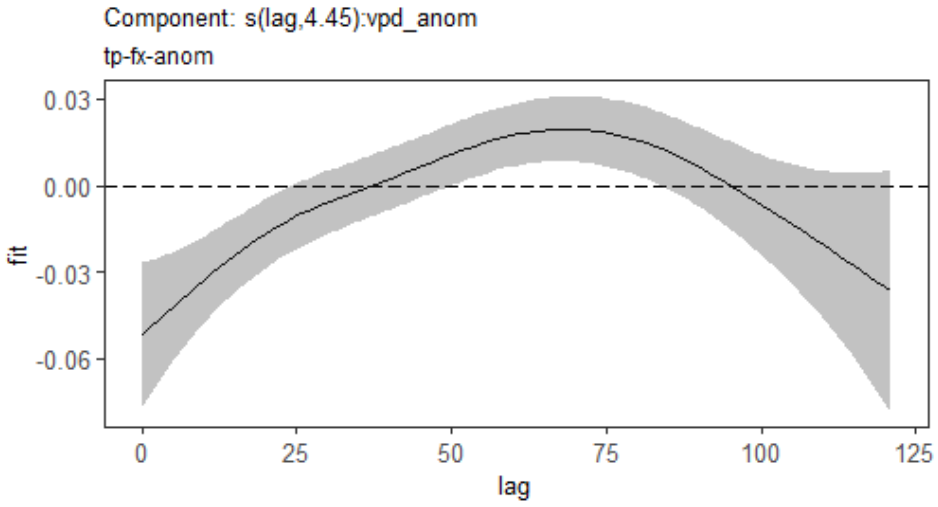

Figure 29: Example 1-D partial effect graph for the smoothed term for the anomalized version of vapor pressure deficit for a particular model. The lag is the number of days of lag (to accomodate the delay in effect from environmental conditions to transmission risk), up to the maximum specified in the parameters (121 here).

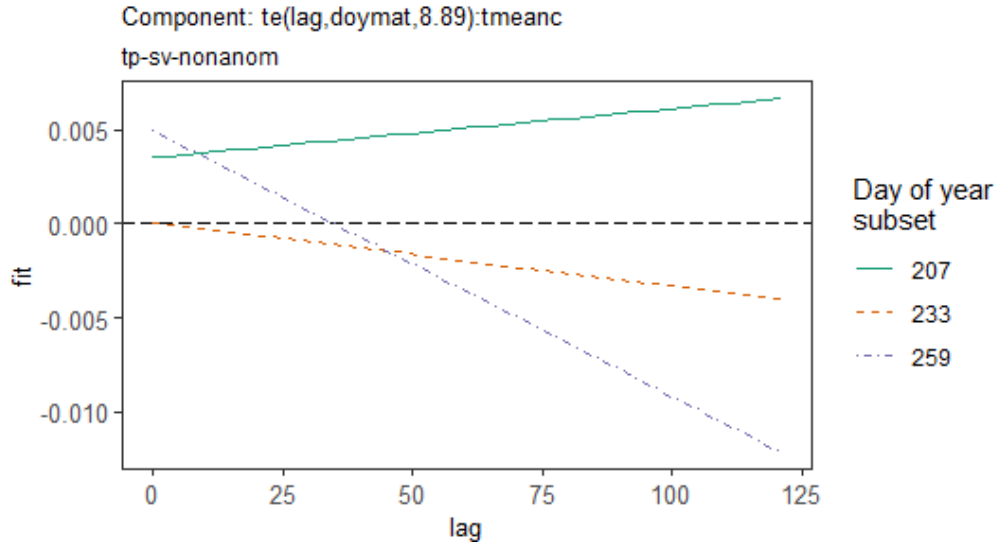

Figure 30: Example 2-D partial effect graph for seasonally-varying mean temperature (observed values) with three sample doys pulled out to graph as lines. The lag is the number of days of lag (to accomodate the delay in effect from environmental conditions to transmission risk), up to the maximum specified in the parameters (121 here).

3.1.7 Reference map

The report will include a list of counties and large state map with all the counties labeled as reference.

3.2 Input data

In the second part of the reports, it will present overviews and summaries of the various input data: human, mosquito, and weather. The narrative will include some summary statistics, e.g. number of rows (cases or pools), confirmation on what years of data were present, and listings of entries that did not match to the spatial data that need manual correction.

3.2.1 Human input data summaries

To provide an overview of the spatial distribution of human WNV cases, ArboMAP will display a county map of the cumulative total of cases in the input data within the year ranges specified by the parameters (Figure 31).

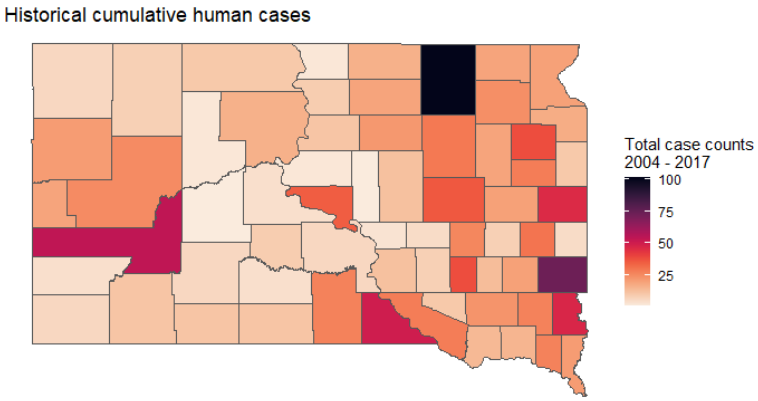

Figure 31: Cumulative human WNV cases map.

For a temporal view, ArboMAP will present a heatmap for the epicurve of human cases in each year (Figure 32). A heatmap is used as it is easier to see when in which year the cases occurred as compared to a many multi-line timeseries chart.

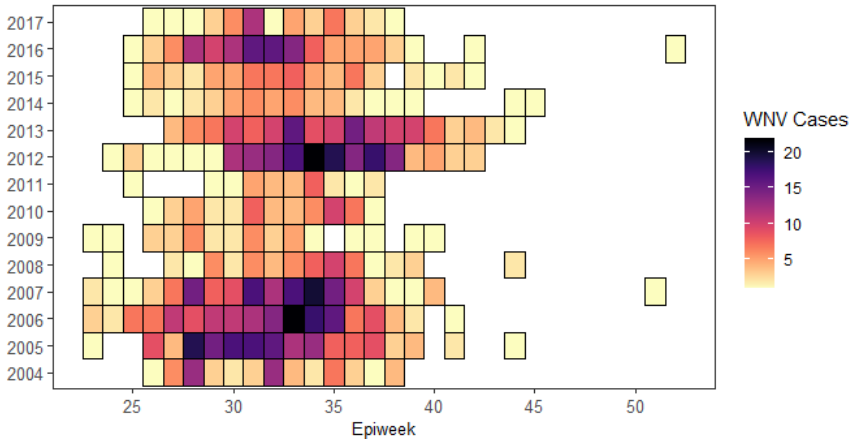

Figure 32: Epicurves as heatmaps for human WNV cases in each year.

3.2.2 Mosquito pool input data

Pool statistics for the past two weeks are also included. If pool data exists for the forecast epiweek, then the two weeks will be the forecast week and the week prior. If data does not exist yet for the requested forecast epiweek, then the weeks shown will be the two epiweeks prior to the forecast week.

A map of the state showing which counties have had positive pools reported in the past two weeks (Figure 33). Counties that did not report pools are marked as such, to distinguish them from counties that had zero positive pools, but did report negative pools.

A table following the map lists counties that have positive pools in the current forecast year to date (and only these counties). Additional columns provide information on the last two weeks as well.

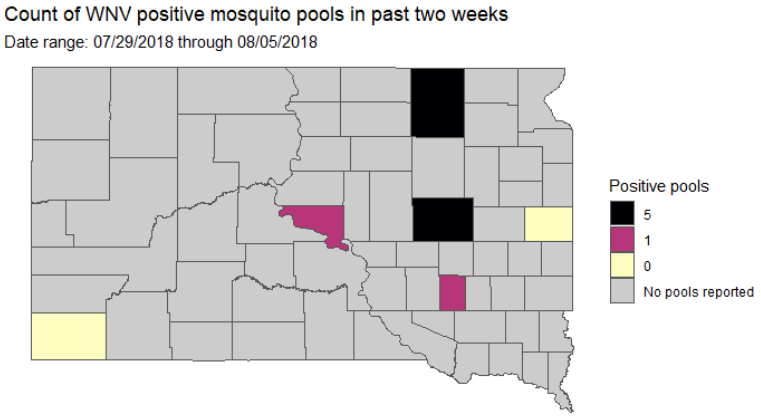

Figure 33: Map of counties that have reported mosquito pool data in the last two weeks.

The next graph shows the percentage of predicted positive pools by year comparing the forecast year red to the requested comparison years (shades of blue) and all other years (gray). If there is sufficient data in the forecast year, the observed pools rates are shown as black dots, binned into six different time points.

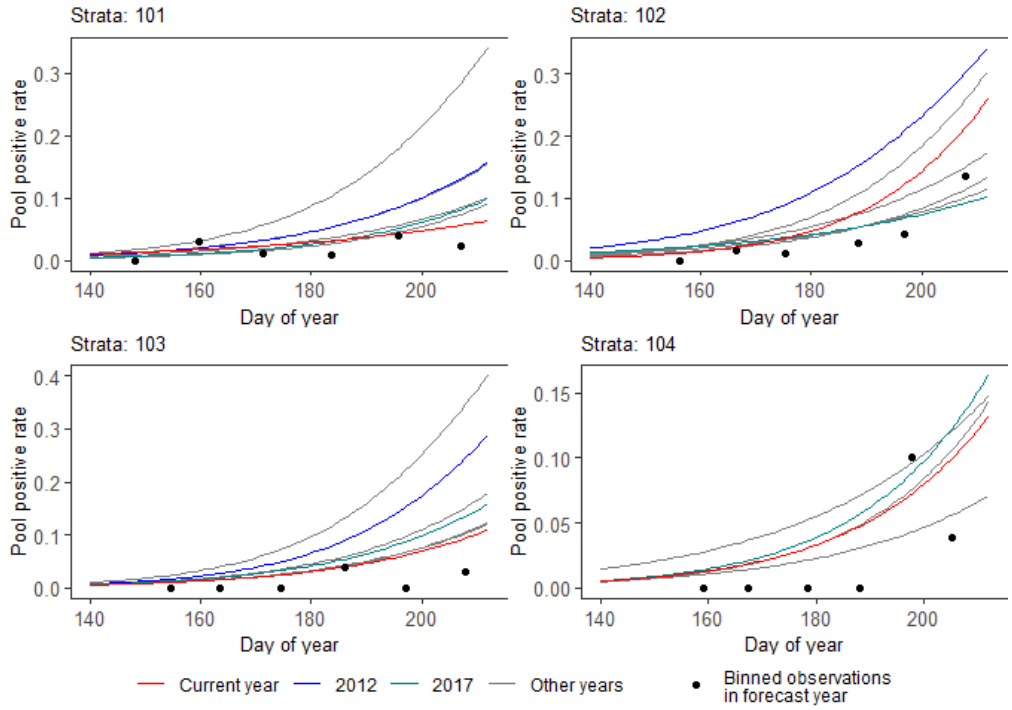

Figure 34: Predicted WNV positive mosquito pools, by year.

The last mosquito graph shows the relative risk due to the mosquito infection rate as a time-series of all known years.

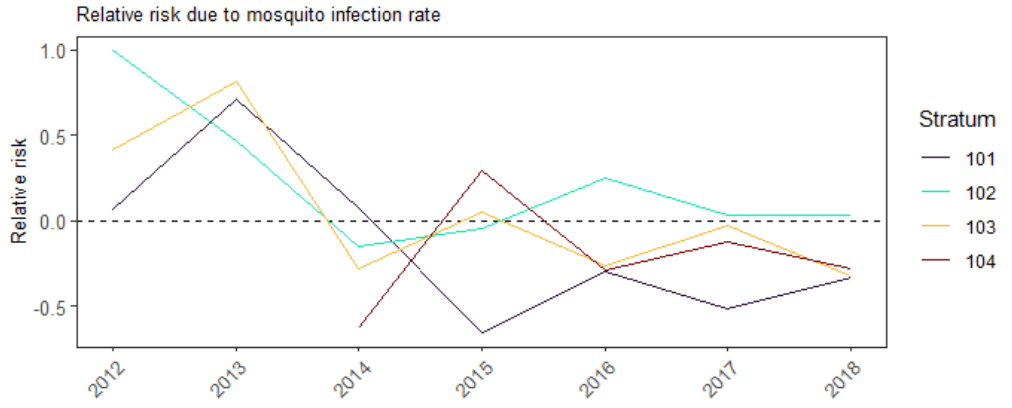

Figure 35: Relative risk from mosquito infection rate.

3.2.3 Weather input data summaries

These graphs show the median state-wide weather variables for the forecast year, compared to the historical median. The main report will have graphs for the *observed* values (Figure 36), and the appendix will show the *anomalized* values (Figure 37).

Two or more consecutive days that are **greater than** the historical median are drawn in red and consecutive days that are **less than** the historical median are drawn in blue. Consecutive days that overlap the historical

median (i.e. one day above and the next below, or the opposite) are in purple. The gray shaded region is a ribbon showing the historical range (min to max).

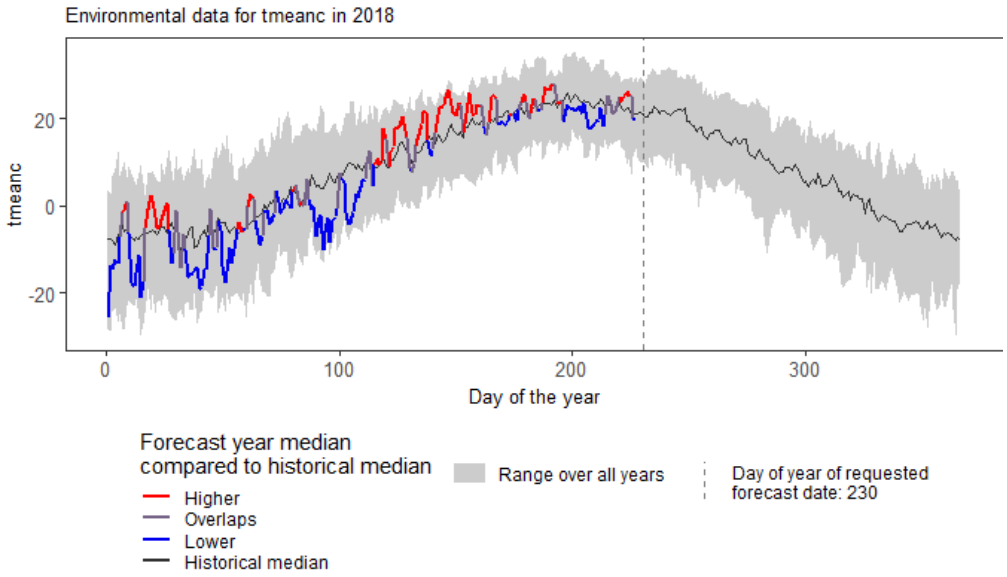

Figure 36: Median state-wide observed weather variables for the current forecast year.

The appendix will show the anomaly graphs, which are the same timeseries with the weather variable has been anomalized (Figure 37). Anomalies are calculated using deviance between the observed value and the predicted value from a GAM regression model using county and a smooth on day of year (seasonality) and county. An anomaly is the observed minus the predicted.

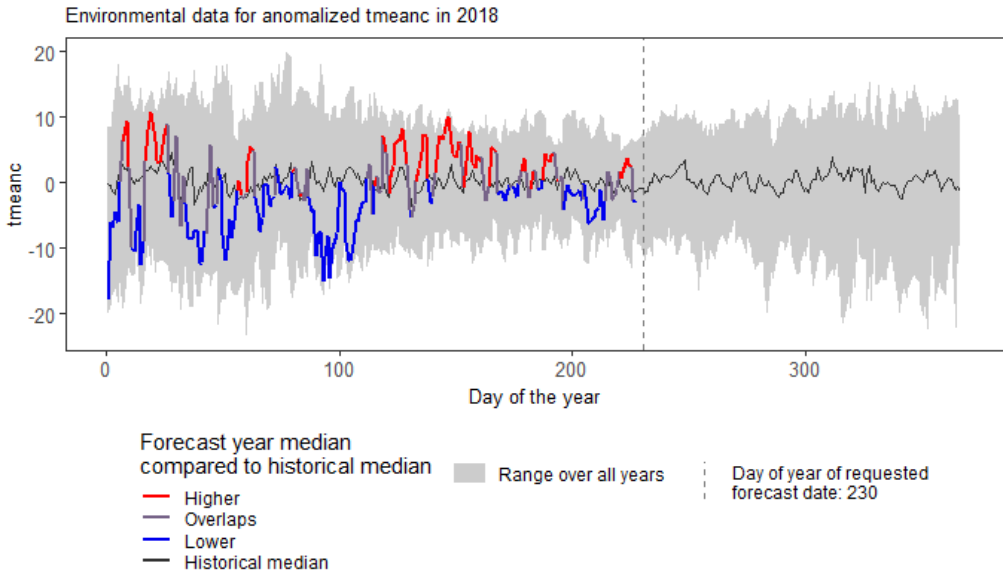

Figure 37: Median state-wide anomalized weather variables for the current forecast year.

3.2.4 Parameters used

At the end of the main report, a chart of the parameters used for this run of the report is included for reference.

4 Relevant scientific papers

- Chuang, T. M. B. Hildreth, M. B., D. L. VanRoekel, and M. C. Wimberly. 2011. Weather and land cover influences on mosquito populations in Sioux Falls, South Dakota. Journal of Medical Entomology 48: 669-679.
- Chuang, T., G. M Henebry, J. S. Kimball, D.L. VanRoekel-Patton, M. B. Hildreth, and M. C. Wimberly. 2012a. Satellite microwave remote sensing for environment modeling of mosquito population dynamics. Remote Sensing of Environment 125: 147-156.
- Chuang, T., C. W. Hockett, L. Kightlinger, and M. C. Wimberly. 2012b. Landscape-level spatial patterns of West Nile virus risk in the northern Great Plains. American Journal of Tropical Medicine and Hygiene 86: 724-731.
- Chuang T., and M. C. Wimberly. 2012. Remote Sensing of Climatic Anomalies and West Nile Virus
- Davis, J. K., G. P. Vincent, M. B. Hildreth, L. Kightlinger, and M. C. Wimberly. 2018. Improving the prediction of arbovirus outbreaks: a comparison of climate-driven models for West Nile virus in an endemic region of the United States. Acta Tropica 185: 242-250.
- Davis J. K., Vincent G. P., Hildreth M. B., Kightlinger L., Carlson C., and M. C. Wimberly. 2017. Integrating Environmental Monitoring and Mosquito Surveillance to Predict Vector-borne Disease: Prospective Forecasts of a West Nile Virus Outbreak. PLoS Currents Outbreaks. 2017 May 23. Edition 1. doi: 10.1371/currents.outbreaks.90e80717c4e67e1a830f17feeaaf85de.
- Hess, A., J. K. Davis, and M. C. Wimberly. 2018. Identifying environmental risk factors and mapping the distribution of West Nile virus in an endemic region of North America. GeoHealth 2:
- Wimberly, M. C., M. B. Hildreth, S. P. Boyte, E. Lindquist, and L. Kightlinger. 2008. Ecological niche of the 2003 West Nile virus epidemic in the northern Great Plains of the United States. PLoS One 3: e3744.
- Wimberly, M. C., P. Giacomo, L. Kightlinger, and M. B. Hildreth. 2013. Spatio-temporal epidemiology of human West Nile virus disease in South Dakota. International Journal of Environmental Research and Public Health 10: 5584-5602.
- Wimberly, M. C., A. Lamsal, P. Giacomo, and T. Chuang. 2014. Regional variation of climatic influences on West Nile virus outbreaks in the United States. American Journal of Tropical Medicine and Hygiene 91: 677-684.