SapFlower

Jiaxin Wang

Automating sap flow data preprocessing, cleaning, modeling, gap-filling, and analysis.

For any inquires and questions,
please contact: Jiaxin Wang (jiaxin.wang@vt.edu; jiaxinwang362@gmail.com)

1. What is SapFlower?

SapFlower is a tool developed for ecologists and plant scientists who are interested in measuring plant sap flow using Thermal Dissipation Probes (TDPs) based on Granier's equation (Granier 1987). Considering measuring sap flow in the field using solar panels powered sap flow sensors may encounter power outage due to the weather conditions; gap-filling is essential for getting complete data to estimate whole growing season water use of interested plants. SapFlower was designed to automate sap flow data cleaning, filtering, segmentation, gap-fill model training, gap-filling, and analysis. Users can feed SapFlower raw data measured from the datalogger, and get complete gap-filled data including dT, K, and F for calculating sap flow.

SapFlower integrates some essential functions such as baseline calculation from the Baseliner 4.

2. How does SapFlower work?

SapFlower has four main parts, namely Project Configuration, Data Preprocessing, Model Training, and Gap-filling, which enable users to handle all sap flow data processing and calculations at one place.

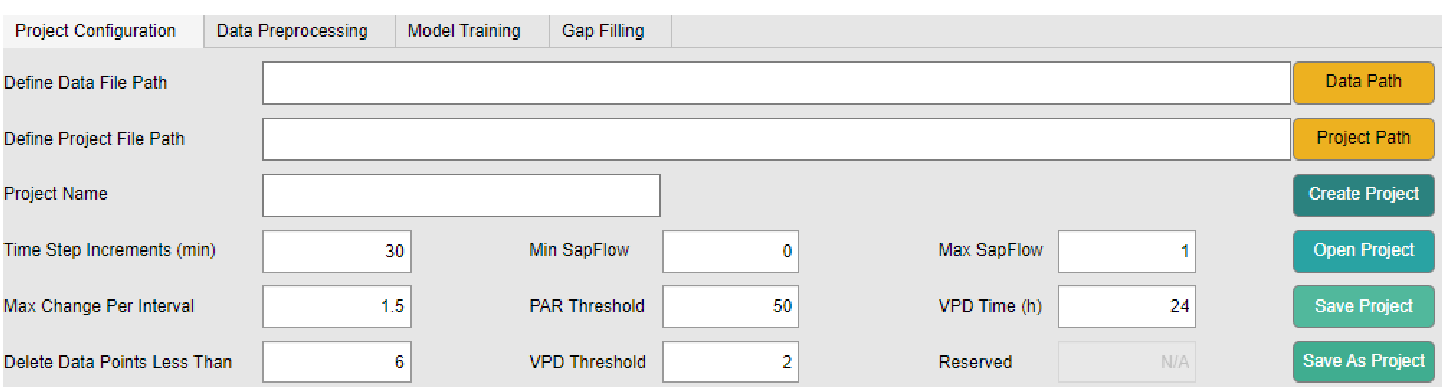
2.1 Project Configuration

Users can create, open, set up parameters for data filtering, and save projects easily through the Project Configuration Tab.

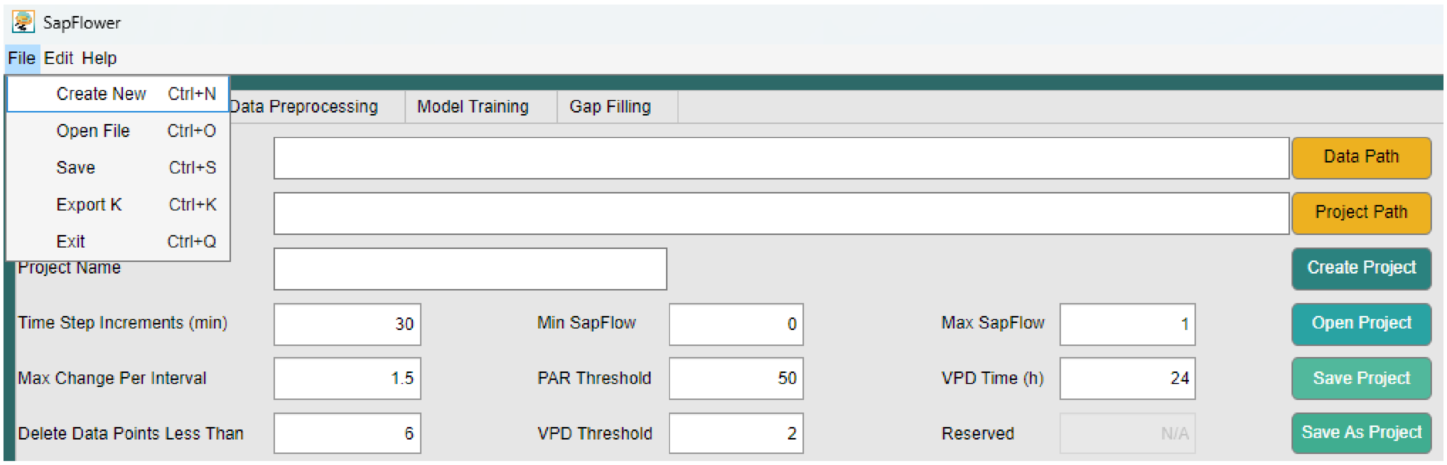
2.1.1 Create a new project

Before creating a project, users can edit the data filtering thresholds for the project. The default values are set to the Toy dataset, users should change them as needed based on their own data.

Field name	***Description***
Define Data File Path	Users can either type or click the “Data Path” button to prompt wot select a data file to edit. The file can be either .csv or .xlsx format, and the file must include some columns for calculation. The mandatory columns are TIMESTAMP, PAR_Den_Avg, AirTC_Avg, RH, and at least one sap flow data column. (Future updates may dynamically handle different users’ data structure). The mandatory columns’ names must be the same as listed here for indexing. Users can also have other columns, such as LAI (leaf area index), DOY, VPD, and it will calculate them based on TIMESTAMP for DOY, and AirTC_Avg and RH for VPD.
Define Project File Path	Users don’t need to define through this edit field, instead, it will prompt to allow user to select a folder to save the project file.
Time Step Increments (min)	Time between two data points. 30 = 30 minutes.
Min SapFlow	Filters data that are below acceptable range. Any dT data less than this value will be deleted. Just set it to zero if you don’t know.
Max SapFlow	Filters data that are above acceptable range. Any dT data greater than this value will be deleted. You can first set to a high value to view your data, and then come back to set it to a reasonable value for your data.
Max Change Per Interval	Filters data spikes. Any value with an absolute change from the previous time step that is greater than this value will be deleted. You can first set to a high value to view your data, and then come back to set it to a reasonable value for your data.
PAR Threshold	Threshold for determining nighttime hours. Note that due to PAR/radiation sensor calibration/drift issues or light pollution, sensors may read some nominal value greater than zero during the night.
VPD Time (h)	Time threshold for low-VPD conditions below which transpiration is expected to cease.
Delete Data Points Less Than	Filters short runs. Due to the outage of battery or low battery, there will be some short runs of measurement, such as only a few of points measured during the noon. You can set it to zero to preview your data first, and then come back to change it to a reasonable value.
VPD Threshold	I would keep this as intact described in Baseliner: Threshold for low-VPD conditions below which transpiration is expected to cease. Note that due to temperature/RH sensor calibration/drift issues, sensors may read some nominal value greater than zero during saturated conditions.



Users can define the data path first, add a name in "Project Name" edit field, and then either use the menu, shortcut (Ctrl+N), or "Create Project" button to prompt to create a project file (.html). After creating a project file, it will save it and open it as the current project file. It will prompt you to select additional environmental or meta data columns. You should select all columns that are not sap flow sensors, so that later you can loop through all your sensors data at once. It may also ask you to select which one is Air temperature and which one is relative humidity if your data doesn't have a VPD column.



2.1.2 Open an existing project file

Users can easily open an existing project file by clicking "Open Project" either using the menu or button or using shortcut (Ctrl+O). It will prompt you to select additional environmental or meta data columns. You should select all columns that are not sap flow sensors, so that later you can loop through all your sensors data at once. It may also ask you to select which one is Air temperature and which one is relative humidity if your data doesn't have a VPD column.

2.1.3 Saving a project file

Users can modify their currently loaded project file and save it as new projects or just update some parameters by clicking "Save Project" either using the menu or button or using shortcut (Ctrl+S).

2.2 Data Preprocessing

This panel is designed to help users to clean their data, and the main design was inspired by Baseliner 4. There are two main options, namely manually clean and auto clean. Manually clean requires users to determine valid and effective data through setting filtering thresholds in Project Configuration and manually selecting and deleting or reversing data points in Data Preprocessing. During data cleaning, if users want to and changed update project configuration, they can click "Plot Data" button to view the updated data based on the new project configuration.

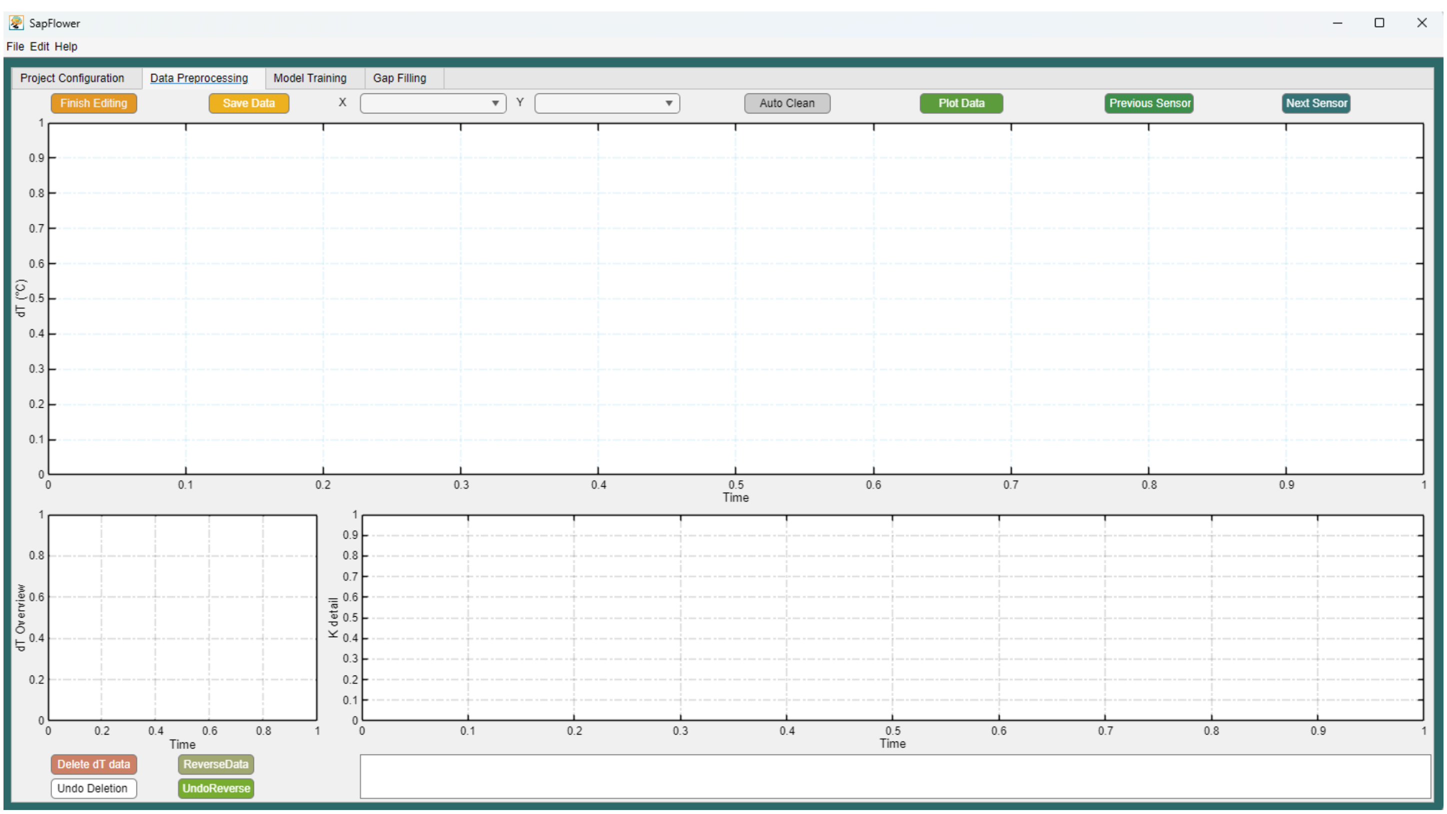
Auto clean requires users to set the auto clean parameters in Model Training left top (WindowSize, IQRThresholdMultiplier, HighVariationThreshold, and LowVariationThreshold). Users can preview by clicking "Auto Clean" to see which data points will be removed by currently defined thresholds. Please note that the Auto Clean will always be applied before all model training to ensure that at least some outliers are removed. It's better if users can manually clean some data and fine-tune the auto clean thresholds to get high quality training data.

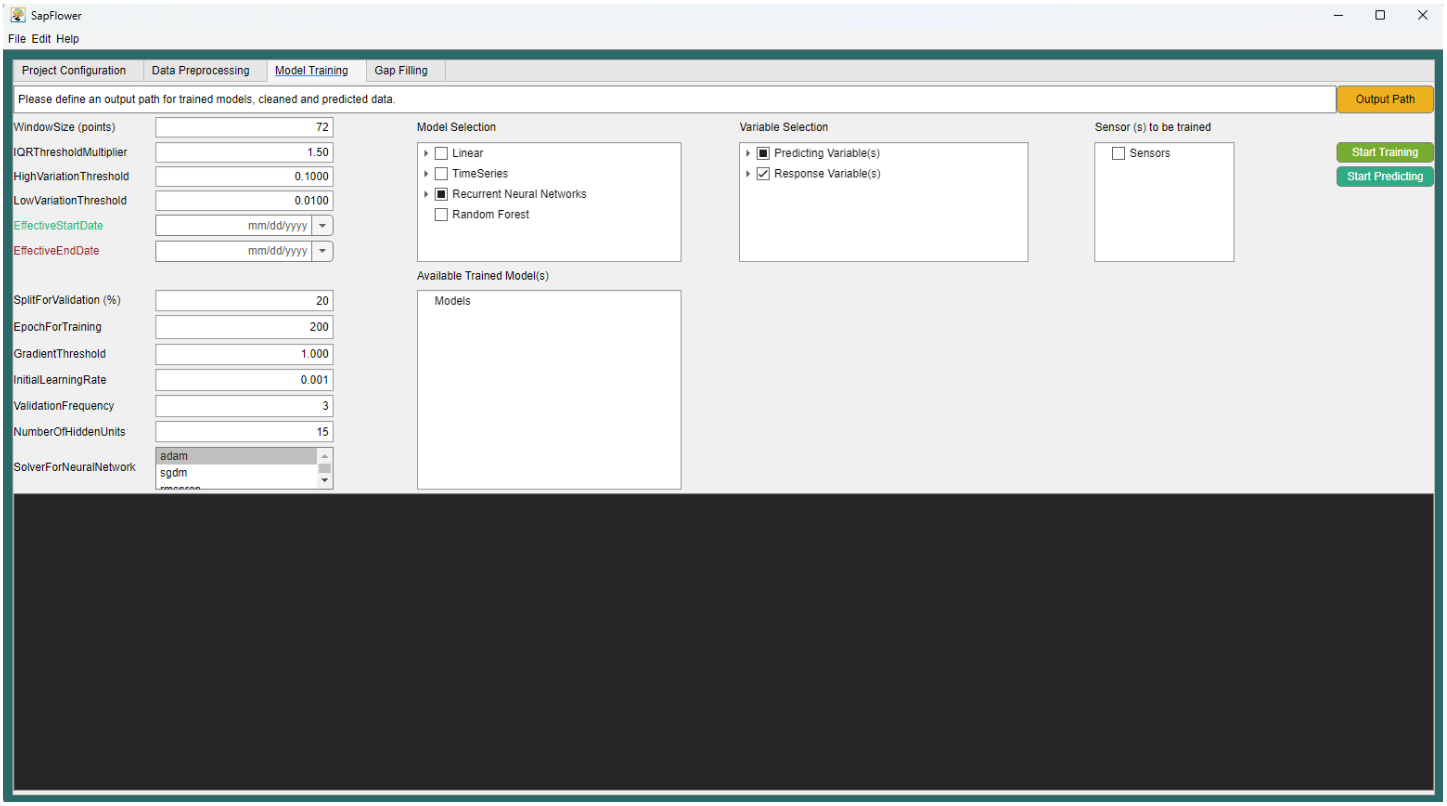
Field name	*Description*
WindowSize	It separates the whole data sequence into multiple short sequences to calculate the quantile, low and high variation. You can set it based on your time step increments. If your time step increments are 30 min, then you can put 72 to represent 36 hours window. Smaller numbers will make it more sensitive.
IQRThresholdMultiplier	Multiplier for IQR to define outliers (Recommend setting it based on your Max SapFlow).
HighVariationThreshold	Threshold for detecting high variation windows. A smaller number will filter more points. (Recommend setting it based on your Max SapFlow).
LowVariationThreshold	Threshold for detecting low variation windows. A larger number will filter more points. (Set a small number like 0.01~0.05 based on your data).

Users can also use shortcuts for data deletion (Ctrl+D), reversal (Ctrl+R), and undo (Ctrl+Z). To switch different sensors, users can select the sensors from the dropdown menu (Y) or just loop through sensors by clicking "Previous Sensor" or "Next Sensor."

Users can also edit the current plot by clicking and editing the title, axes, and legend, or export plot by moving the mouse to the right corner of the plot to use the export plot tool. The plot can be exported in .JPG, .PNG, .TIF, and .PDF format. Users may see that there are some data smoothing functions when right click on the plot, but please note that those haven't been enabled to use. Future updates may enable those functions. All executing output will be updated in the bottom text area, and users can use it as references when they encounter any unexpected errors or warnings.

Once users finished cleaning, they can go to the Model Training Tab without saving the edited data since it will be updated automatically to the current data table. However, if users want to save the cleaned data, they can then click "Save Data" to save the cleaned data and export the current K values. Alternatively, users can save to replace their original data or save as a new data file by clicking "Finish Editing." The time for saving data depends on the data that users loaded, so please wait until it finishes saving.





2.3 Model Training

This Tab is designed to help users train their own models for sap flow data gap-filling. Before model training, users need to set their output data path for cleaned data, trained model, and predicted and gap-filled data. Users just need to set one path; it will set sub folders to store different types of data.

There are several steps that users should confirm before clicking the "Start Training" button, and the "Start Prediction" button will only work after model training.

Step 1. Set output path

Users need to ensure that their data is cleaned/or set up reasonable auto clean parameters for auto cleaning.

Step 2. Set parameters for Auto Clean

Define effective growth period through the Date pickers. The gap-filled output data will use them as references for calculations.

Step 3. Model Selection

Users may select/check at least and only one model for gap-filling model training. Please note that Time Series models (ARX and ARMAX) require Y values as reference for prediction and can only be used for short period (e.g., 24 hours) predictions. If users have more over 24 hours missing period, I would recommend using Recurrent Neural Networks or Random Forest models.

Step 4. Variable Selection

In the current version, users only need to select the Predicting Variables for model training. The variables will be listed under Predicting Variable(s) Tree once the users loaded their project. Users can decide to check one or multiple variables for different models of training. Here VPD is already checked as default considering their substantial relationship with sap flow. Users may leave the Response Variable(s) as default, and they can export K and F values once they finished gap-filling.

Step 5. Sensor(s) to be trained

Sensors will be available for checking once users loaded their project. Users may choose multiple sensors at once to loop through and train models for all checked sensors.

Step 6. Model training hyper parameters

Field name	***Description***
SplitForValidation (%)	This is where users can set their data splitting for model training and validation. A number like 20 represents that 20% data will be split to be used for model validation. Users can try different numbers to see how it will affect the model training.
EpochForTraining	Users can determine the numbers of iterations for model training, and this will be only applied to Recurrent Neural Networks. A larger number may get more well trained or over trained model depending on your data and variables selected for model training. You can leave it as default if you don't know how to set.
GradientThreshold	Here it is used to prevent gradient explosion: In deep networks, especially recurrent networks, gradients can sometimes grow uncontrollably large, leading to instability. Clipping helps keep them in check. Improved stability: It improves training stability by ensuring that excessively large gradient updates don't destabilize the optimization process.
InitialLearningRate	The learning rate determines the size of the steps the optimization algorithm takes to minimize the loss function during training. If the learning rate is too high, the model may overshoot the minimum of the loss function, leading to instability. If the learning rate is too low, training may be very slow or get stuck in a local minimum.
ValidationFrequency	This parameter is used for validation. If you set it to one, it will validate immediately after each iteration of training, and if you set it to 5, it will validate every 5 iterations of training.
NumberOfHiddenUnits	The hidden units represent the neurons in the hidden layer that learn internal representations of the input data. The number of hidden units influences the model's ability to capture complex patterns in the data. More hidden units can increase the capacity of the model to learn more intricate relationships, but it also increases the computational complexity and the risk of overfitting if the dataset is small.
SolverForNeuralNetwork	It refers to the optimization algorithm (solver) used to minimize the loss function and update the model weights during training. The solver plays a critical role in determining how the neural network learns and converges. (Continued on next page)

Field name	*Description*
SolverForNeuralNetwork	'adam' (Adaptive Moment Estimation): This is a popular solver that combines the benefits of RMSProp and momentum. It adjusts the learning rate for each parameter and can perform well in many scenarios, especially for deep networks and LSTM networks. It's often a good starting choice for deep learning models because of its adaptive learning rate feature. 'sgdm' (Stochastic Gradient Descent with Momentum): This is the standard stochastic gradient descent method with momentum to accelerate convergence and reduce oscillations. It's effective for many types of neural networks, including CNNs and fully connected networks. 'rmsprop' (Root Mean Square Propagation): This solver is designed to deal with the diminishing learning rates problem in the traditional stochastic gradient descent method by normalizing the gradients based on their recent magnitudes. It is especially useful for recurrent neural networks and deep networks.

Step 7. Start Training.

Once users clicked to start the training process, the data will be auto cleaned and saved to CleanedData folder, and then the model training process will start. Users will be able to monitor the model training process either via the text output or real-time plotting (due to MATLAB license limit, the real-time plotting will be only available for those who are using SapFlower in MATLAB). Trained models will be saved to TrainedModels folder.

Step 8. Start Predicting.

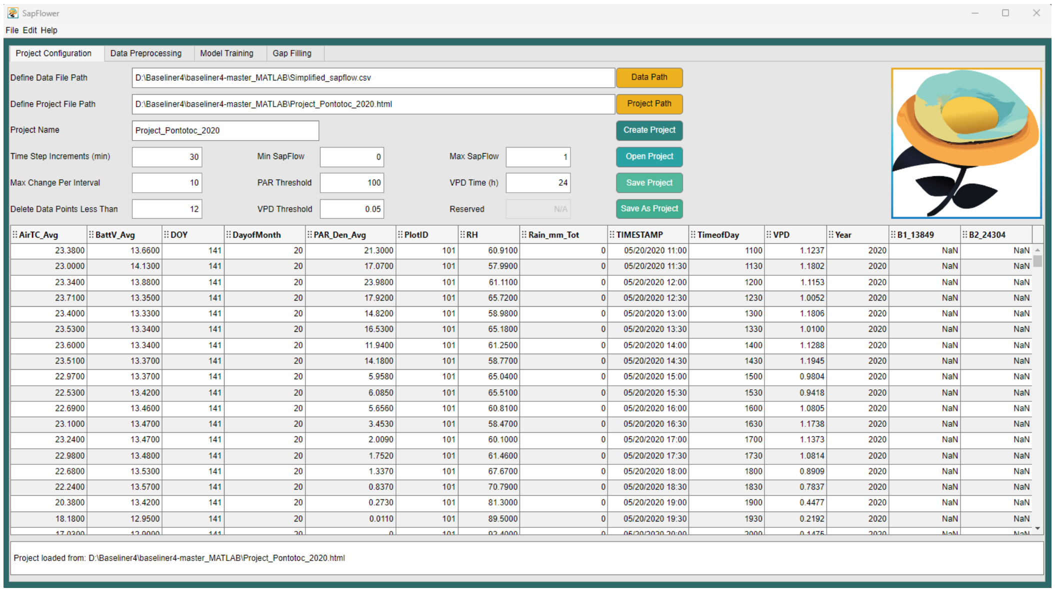
Once users finish model training, they can start model predicting, and the predicted data for each sensor with model types will be saved to PredictedData folder.

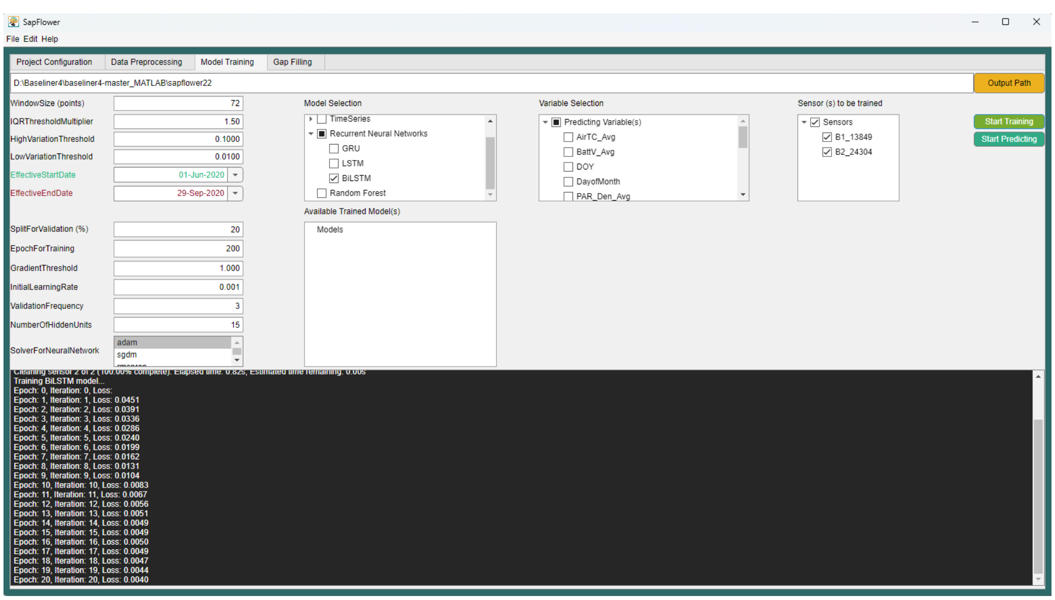
2.4 Gap-filling

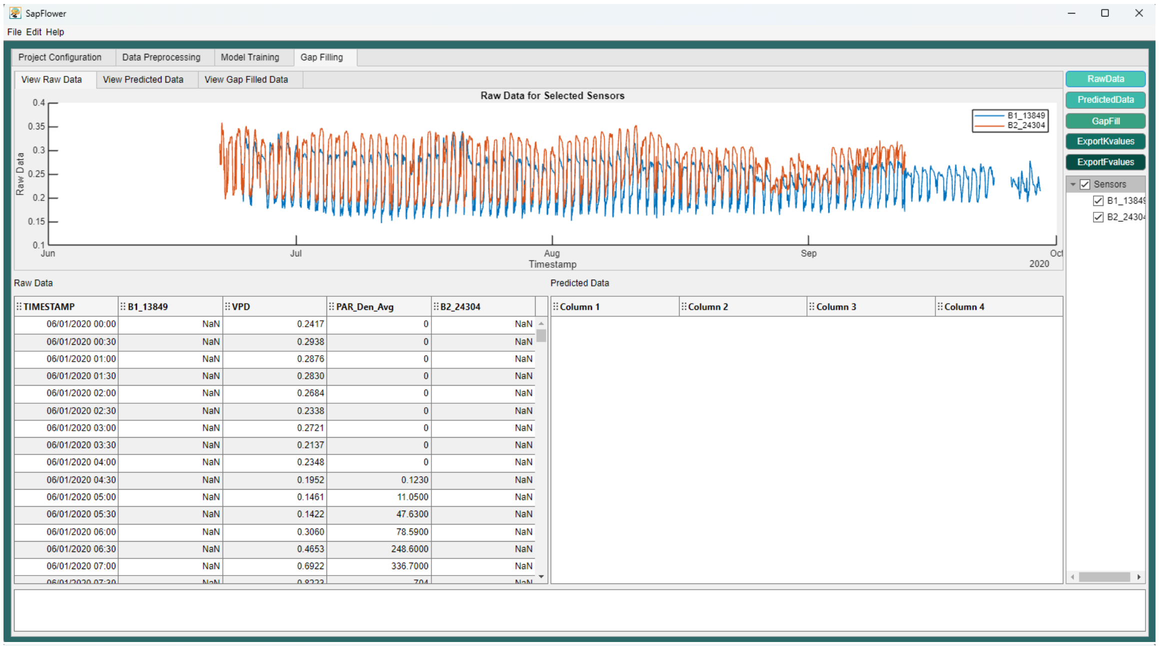
This Tab is designed to help users do and visualize gap-filling. Once users finish model prediction, the sensors available for gap-filling will be listed under the right list tree. To do all operations, users must check at least one sensor to be gap-filled. Users can start loading and view cleaned raw data in both table and figure by clicking the "RawData" button. Similarly, users can click the View Predicted Data Tab to view the predicted data plot by clicking "PredictedData" button. After loading both raw data and predicted data, users can then do gap-fill by clicking the "GapFill" button. Gap-filled data will be saved to Gapilled folder. Users can also edit, zoom in/out, and export all plots through using the plot's right corner's plotting tool bar.

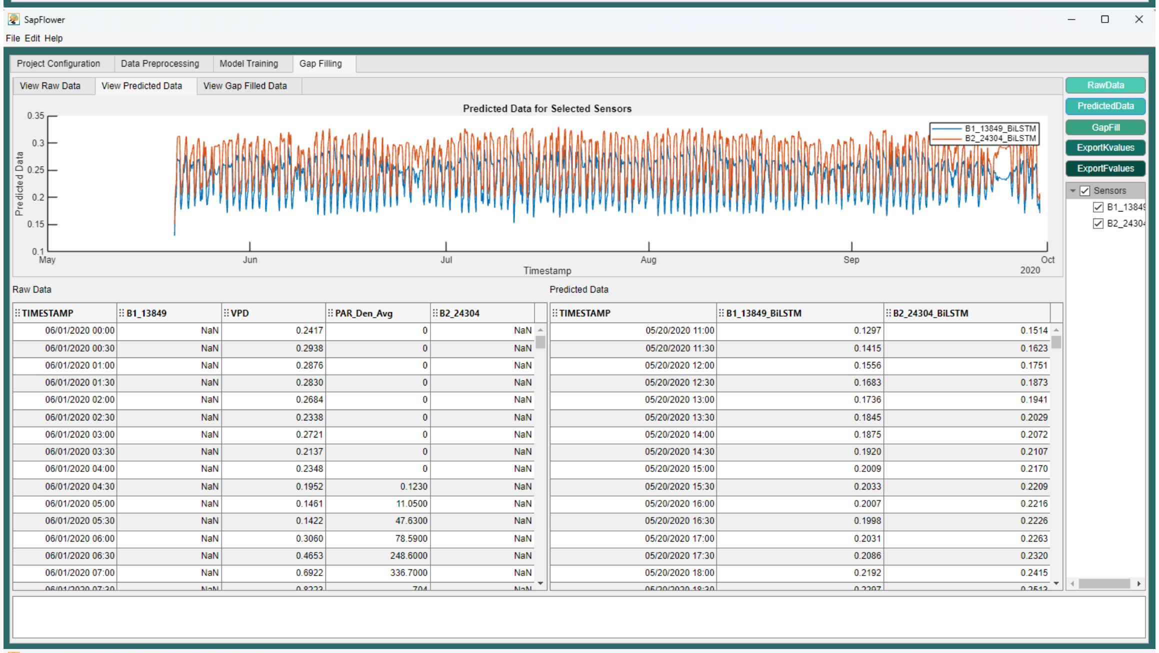
Once users finished gap-filling, they can export K and/or F values through clicking "ExportKvalues" and/or "ExportFvalues", and the exported data will be saved to ExportedKvalue and/or ExportedFvalue folder. Users can then do further analysis using that data.

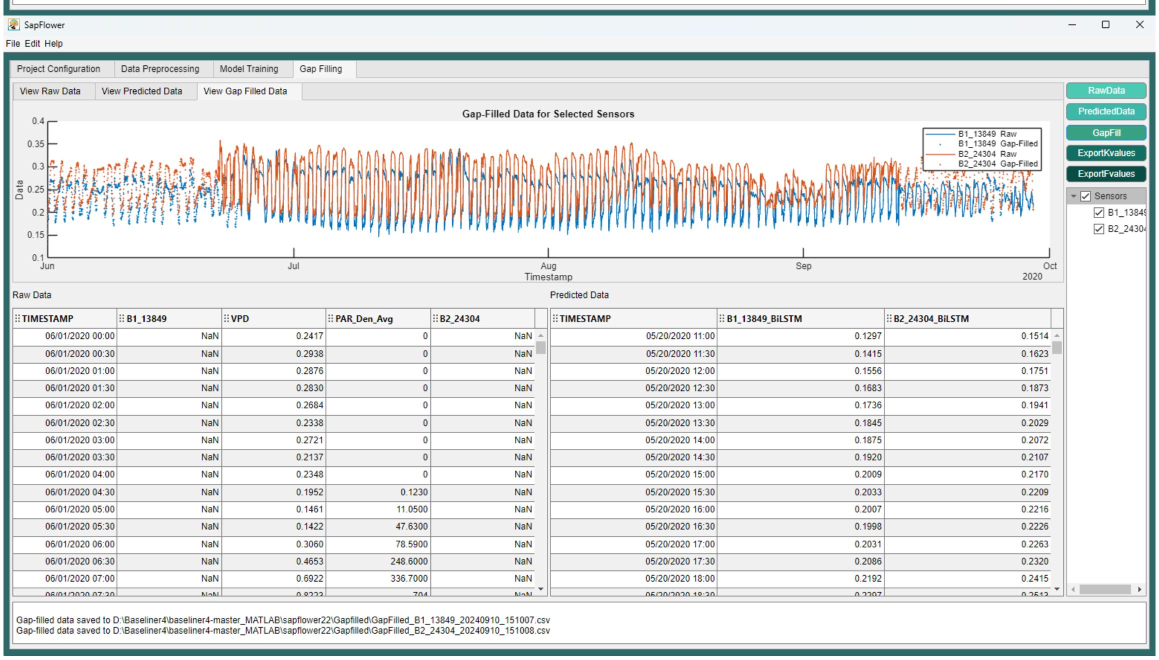
Considering the sapwood area and tree growing patterns varied case by case, daily water use calculation, and some other analyses are currently not available in SapFlower. Future updates may include those high-level calculations.



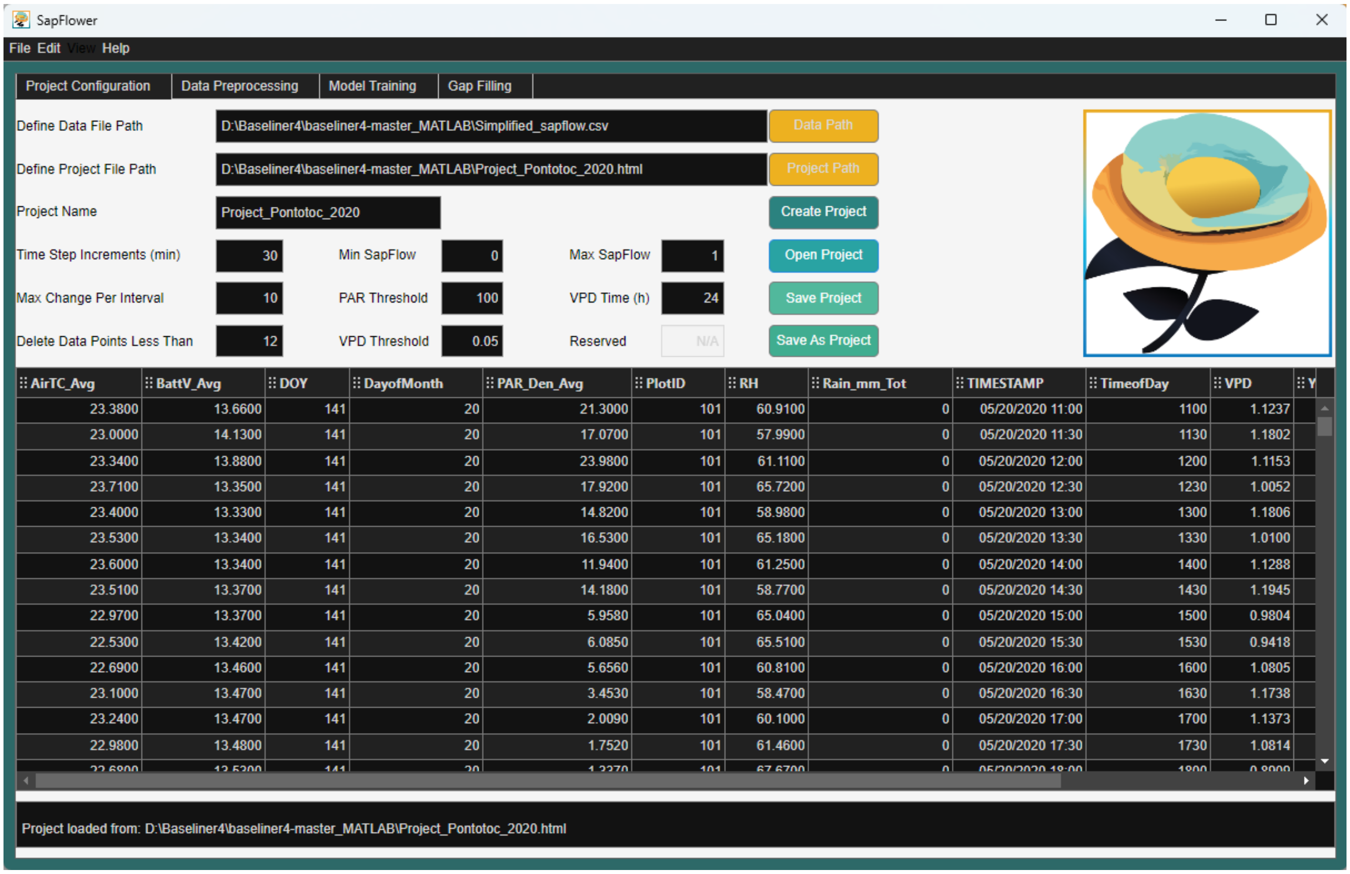


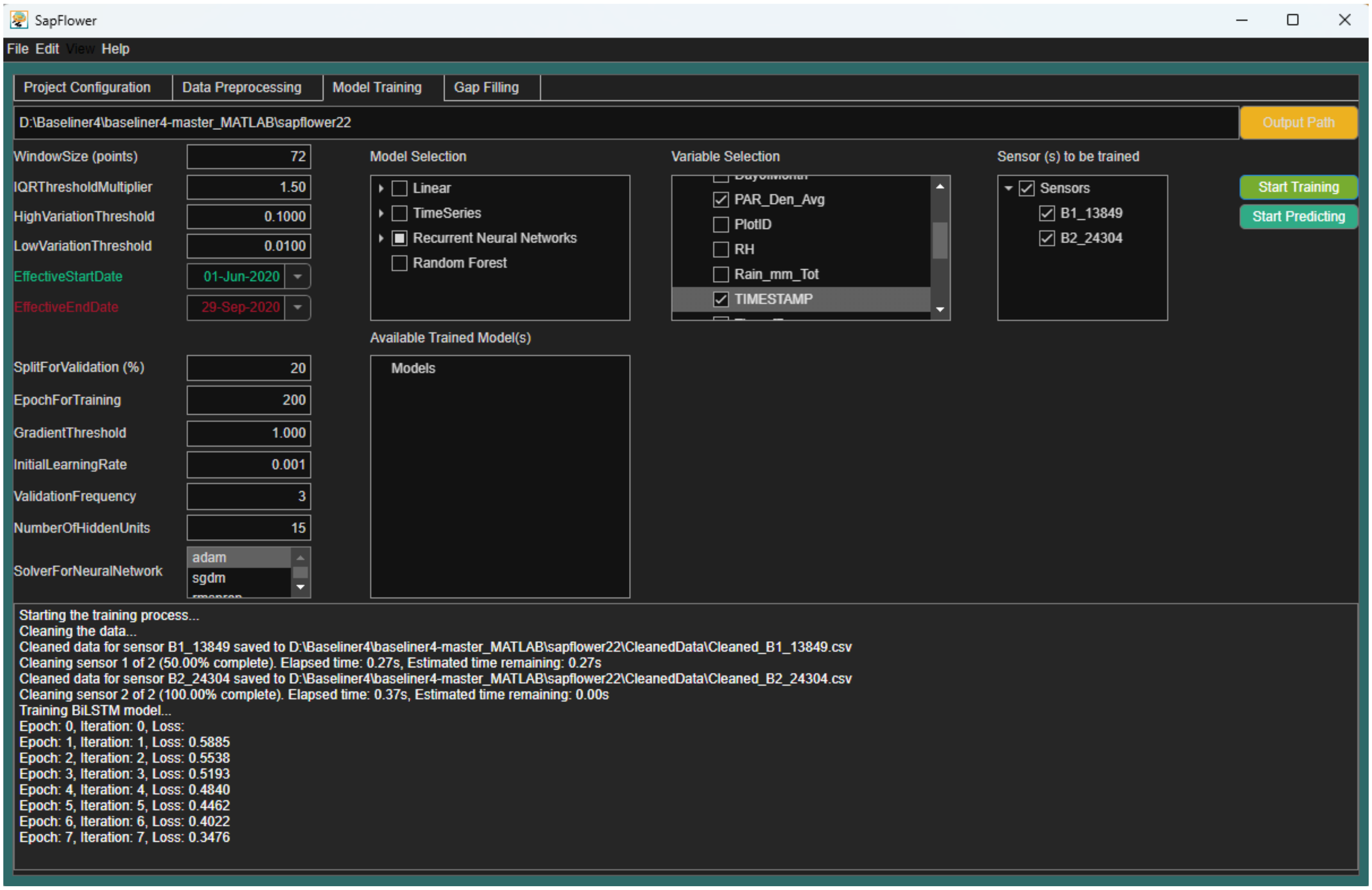

SapFlower
File Edit Help
Project Configuration
Data Preprocessing
Model Training
Gap Filling
View Raw Data
View Predicted Data
View Gap Filled Data
Raw Data for Selected Sensors
B1_13849
B2_24304
Raw Data
Timestamp
Jun
Jul
Aug
Sep
Oct
2020
RawData
PredictedData
GapFill
ExportKvalues
ExportFvalues
Sensors
Raw Data
Predicted Data
TIMESTAMP B1_13849 VPD PAR_Den_Avg B2_24304
06/01/2020 00:00 NaN 0.2417 0 NaN
06/01/2020 00:30 NaN 0.2938 0 NaN
06/01/2020 01:00 NaN 0.2876 0 NaN
06/01/2020 01:30 NaN 0.2830 0 NaN
06/01/2020 02:00 NaN 0.2684 0 NaN
06/01/2020 02:30 NaN 0.2338 0 NaN
06/01/2020 03:00 NaN 0.2721 0 NaN
06/01/2020 03:30 NaN 0.2137 0 NaN
06/01/2020 04:00 NaN 0.2348 0 NaN
06/01/2020 04:30 NaN 0.1952 0.1230 NaN
06/01/2020 05:00 NaN 0.1461 11.0500 NaN
06/01/2020 05:30 NaN 0.1422 47.6300 NaN
06/01/2020 06:00 NaN 0.3060 78.5900 NaN
06/01/2020 06:30 NaN 0.4653 248.6000 NaN
06/01/2020 07:00 NaN 0.6922 336.7000 NaN
Column 1
Column 2
Column 3
Column 4

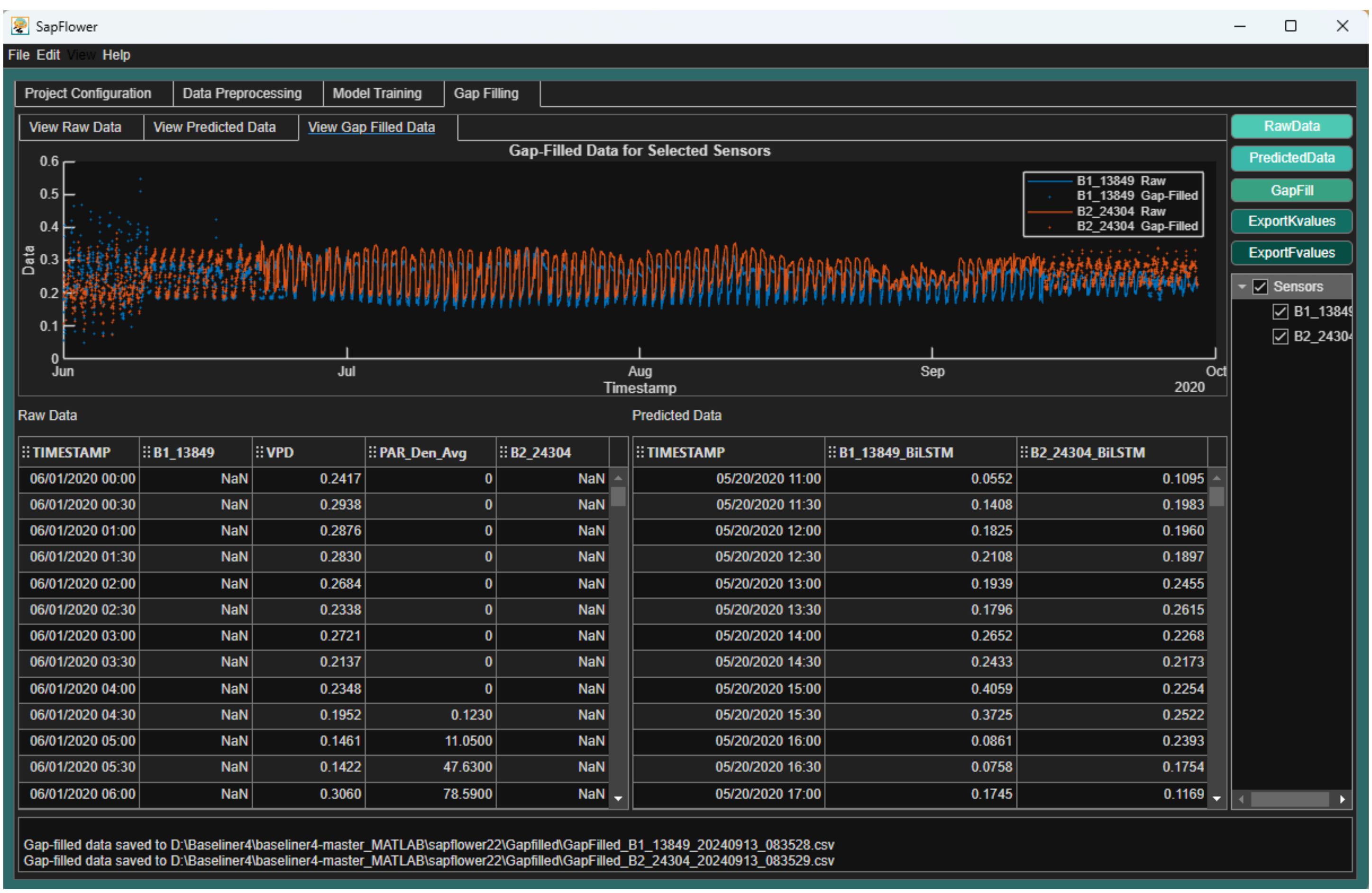

SapFlower
File Edit Help
Project Configuration
Data Preprocessing
Model Training
Gap Filling
View Raw Data
View Predicted Data
View Gap Filled Data
Predicted Data for Selected Sensors
B1_13849_BiLSTM
B2_24304_BiLSTM
Predicted Data
Timestamp
May
Jun
Jul
Aug
Sep
Oct
2020
RawData
PredictedData
GapFill
ExportKvalues
ExportFvalues
Sensors
Raw Data
Predicted Data
TIMESTAMP B1_13849 VPD PAR_Den_Avg B2_24304
06/01/2020 00:00 NaN 0.2417 0 NaN
06/01/2020 00:30 NaN 0.2938 0 NaN
06/01/2020 01:00 NaN 0.2876 0 NaN
06/01/2020 01:30 NaN 0.2830 0 NaN
06/01/2020 02:00 NaN 0.2684 0 NaN
06/01/2020 02:30 NaN 0.2338 0 NaN
06/01/2020 03:00 NaN 0.2721 0 NaN
06/01/2020 03:30 NaN 0.2137 0 NaN
06/01/2020 04:00 NaN 0.2348 0 NaN
06/01/2020 04:30 NaN 0.1952 0.1230 NaN
06/01/2020 05:00 NaN 0.1461 11.0500 NaN
06/01/2020 05:30 NaN 0.1422 47.6300 NaN
06/01/2020 06:00 NaN 0.3060 78.5900 NaN
06/01/2020 06:30 NaN 0.4653 248.6000 NaN
06/01/2020 07:00 NaN 0.6922 336.7000 NaN
TIMESTAMP B1_13849_BiLSTM B2_24304_BiLSTM
05/20/2020 11:00 0.1297 0.1514
05/20/2020 11:30 0.1415 0.1623
05/20/2020 12:00 0.1556 0.1751
05/20/2020 12:30 0.1683 0.1873
05/20/2020 13:00 0.1736 0.1941
05/20/2020 13:30 0.1845 0.2029
05/20/2020 14:00 0.1875 0.2072
05/20/2020 14:30 0.1920 0.2107
05/20/2020 15:00 0.2009 0.2170
05/20/2020 15:30 0.2033 0.2209
05/20/2020 16:00 0.2007 0.2216
05/20/2020 16:30 0.1998 0.2226
05/20/2020 17:00 0.2031 0.2263
05/20/2020 17:30 0.2086 0.2320
05/20/2020 18:00 0.2192 0.2415


SapFlower
File Edit Help
Project Configuration
Data Preprocessing
Model Training
Gap Filling
View Raw Data
View Predicted Data
View Gap Filled Data
Gap-Filled Data for Selected Sensors
B1_13849 Raw
B1_13849 Gap-Filled
B2_24304 Raw
B2_24304 Gap-Filled
Data
Timestamp
Jun
Jul
Aug
Sep
Oct
2020
RawData
PredictedData
GapFill
ExportKvalues
ExportFvalues
Sensors
Raw Data
Predicted Data
TIMESTAMP B1_13849 VPD PAR_Den_Avg B2_24304
06/01/2020 00:00 NaN 0.2417 0 NaN
06/01/2020 00:30 NaN 0.2938 0 NaN
06/01/2020 01:00 NaN 0.2876 0 NaN
06/01/2020 01:30 NaN 0.2830 0 NaN
06/01/2020 02:00 NaN 0.2684 0 NaN
06/01/2020 02:30 NaN 0.2338 0 NaN
06/01/2020 03:00 NaN 0.2721 0 NaN
06/01/2020 03:30 NaN 0.2137 0 NaN
06/01/2020 04:00 NaN 0.2348 0 NaN
06/01/2020 04:30 NaN 0.1952 0.1230 NaN
06/01/2020 05:00 NaN 0.1461 11.0500 NaN
06/01/2020 05:30 NaN 0.1422 47.6300 NaN
06/01/2020 06:00 NaN 0.3060 78.5900 NaN
06/01/2020 06:30 NaN 0.4653 248.6000 NaN
06/01/2020 07:00 NaN 0.6922 336.7000 NaN
TIMESTAMP B1_13849_BiLSTM B2_24304_BiLSTM
05/20/2020 11:00 0.1297 0.1514
05/20/2020 11:30 0.1415 0.1623
05/20/2020 12:00 0.1556 0.1751
05/20/2020 12:30 0.1683 0.1873
05/20/2020 13:00 0.1736 0.1941
05/20/2020 13:30 0.1845 0.2029
05/20/2020 14:00 0.1875 0.2072
05/20/2020 14:30 0.1920 0.2107
05/20/2020 15:00 0.2009 0.2170
05/20/2020 15:30 0.2033 0.2209
05/20/2020 16:00 0.2007 0.2216
05/20/2020 16:30 0.1998 0.2226
05/20/2020 17:00 0.2031 0.2263
05/20/2020 17:30 0.2086 0.2320
05/20/2020 18:00 0.2192 0.2415
Gap-filled data saved to D:\Baseliner4\baseliner4-master_MATLAB\sapflower22\Gapfilled\GapFilled_B1_13849_20240910_151007.csv
Gap-filled data saved to D:\Baseliner4\baseliner4-master_MATLAB\sapflower22\Gapfilled\GapFilled_B2_24304_20240910_151008.csv

If you run SapFlower in MATLAB and you have New Desktop for MATLAB (Beta) installed, you will be able to switch Dark theme and Light theme.




SapFlower
File Edit View Help
Project Configuration
Data Preprocessing
Model Training
Gap Filling
D:\Baseliner4\baseliner4-master_MATLAB\sapflower22
Output Path
WindowSize (points) 72
IQRThresholdMultiplier 1.50
HighVariationThreshold 0.1000
LowVariationThreshold 0.0100
EffectiveStartDate 01-Jun-2020
EffectiveEndDate 29-Sep-2020
SplitForValidation (%) 20
EpochForTraining 200
GradientThreshold 1.000
InitialLearningRate 0.001
ValidationFrequency 3
NumberOfHiddenUnits 15
SolverForNeuralNetwork
adam
sgdm
Model Selection
Linear
TimeSeries
Recurrent Neural Networks
Random Forest
Available Trained Model(s)
Models
Variable Selection
PAR_Den_Avg
PlotID
RH
Rain_mm_Tot
TIMESTAMP
Sensor (s) to be trained
Sensors
B1_13849
B2_24304
Start Training
Start Predicting
Starting the training process...
Cleaning the data...
Cleaned data for sensor B1_13849 saved to D:\Baseliner4\baseliner4-master_MATLAB\sapflower22\CleanedData\Cleaned_B1_13849.csv
Cleaning sensor 1 of 2 (50.00% complete). Elapsed time: 0.27s, Estimated time remaining: 0.27s
Cleaned data for sensor B2_24304 saved to D:\Baseliner4\baseliner4-master_MATLAB\sapflower22\CleanedData\Cleaned_B2_24304.csv
Cleaning sensor 2 of 2 (100.00% complete). Elapsed time: 0.37s, Estimated time remaining: 0.00s
Training BiLSTM model...
Epoch: 0, Iteration: 0, Loss:
Epoch: 1, Iteration: 1, Loss: 0.5885
Epoch: 2, Iteration: 2, Loss: 0.5538
Epoch: 3, Iteration: 3, Loss: 0.5193
Epoch: 4, Iteration: 4, Loss: 0.4840
Epoch: 5, Iteration: 5, Loss: 0.4462
Epoch: 6, Iteration: 6, Loss: 0.4022
Epoch: 7, Iteration: 7, Loss: 0.3476


SapFlower
File Edit View Help
Project Configuration
Data Preprocessing
Model Training
Gap Filling
View Raw Data
View Predicted Data
View Gap Filled Data
Gap-Filled Data for Selected Sensors
B1_13849 Raw
B1_13849 Gap-Filled
B2_24304 Raw
B2_24304 Gap-Filled
Data
Timestamp
Jun
Jul
Aug
Sep
Oct
2020
RawData
PredictedData
GapFill
ExportKvalues
ExportFvalues
Sensors
Raw Data
TIMESTAMP B1_13849 VPD PAR_Den_Avg B2_24304
06/01/2020 00:00 NaN 0.2417 0 NaN
06/01/2020 00:30 NaN 0.2938 0 NaN
06/01/2020 01:00 NaN 0.2876 0 NaN
06/01/2020 01:30 NaN 0.2830 0 NaN
06/01/2020 02:00 NaN 0.2684 0 NaN
06/01/2020 02:30 NaN 0.2338 0 NaN
06/01/2020 03:00 NaN 0.2721 0 NaN
06/01/2020 03:30 NaN 0.2137 0 NaN
06/01/2020 04:00 NaN 0.2348 0 NaN
06/01/2020 04:30 NaN 0.1952 0.1230 NaN
06/01/2020 05:00 NaN 0.1461 11.0500 NaN
06/01/2020 05:30 NaN 0.1422 47.6300 NaN
06/01/2020 06:00 NaN 0.3060 78.5900 NaN
Predicted Data
TIMESTAMP B1_13849_BiLSTM B2_24304_BiLSTM
05/20/2020 11:00 0.0552 0.1095
05/20/2020 11:30 0.1408 0.1983
05/20/2020 12:00 0.1825 0.1960
05/20/2020 12:30 0.2108 0.1897
05/20/2020 13:00 0.1939 0.2455
05/20/2020 13:30 0.1796 0.2615
05/20/2020 14:00 0.2652 0.2268
05/20/2020 14:30 0.2433 0.2173
05/20/2020 15:00 0.4059 0.2254
05/20/2020 15:30 0.3725 0.2522
05/20/2020 16:00 0.0861 0.2393
05/20/2020 16:30 0.0758 0.1754
05/20/2020 17:00 0.1745 0.1169
Gap-filled data saved to D:\Baseliner4\baseliner4-master_MATLAB\sapflower22\Gapfilled\GapFilled_B1_13849_20240913_083528.csv
Gap-filled data saved to D:\Baseliner4\baseliner4-master_MATLAB\sapflower22\Gapfilled\GapFilled_B2_24304_20240913_083529.csv