

# User's Manual for Downscaler Fusion Software

Matthew Heaton<sup>1</sup>

David Holland<sup>2</sup>

Thomas Leininger<sup>3</sup>

January 27, 2012

## Contents

<b>1</b>	<b>What You Must Know About Fit_Downscaler.m</b>	<b>2</b>
<b>2</b>	<b>Preliminaries</b>	<b>2</b>
2.1	Overview . . . . .	3
2.2	Methodology . . . . .	5
2.3	Fitting the Model . . . . .	5
2.4	Formatting Your Data . . . . .	6
<b>3</b>	<b>Initial Call to Fit_Downscaler.m - Dialog Boxes</b>	<b>7</b>
3.1	README Dialog Box . . . . .	7
3.2	File Browser . . . . .	7
3.3	MCMC Settings Dialog Box . . . . .	8
3.4	Output Dialog Box . . . . .	8
3.5	Data Format Dialog Boxes . . . . .	9
3.6	Options Dialog Boxes . . . . .	11
<b>4</b>	<b>Options for Fit_Downscaler.m</b>	<b>12</b>
4.1	Validation . . . . .	12
4.2	Prediction . . . . .	13
<b>5</b>	<b>Output from a call to Fit_Downscaler.m</b>	<b>14</b>
<b>6</b>	<b>Troubleshooting</b>	<b>14</b>

---

<sup>1</sup>National Center for Atmospheric Research, Boulder, CO 80307

<sup>2</sup>U.S. Environmental Protection Agency, National Exposure Research Laboratory, RTP, NC 27711

<sup>3</sup>Duke University, Durham, NC 27708

# 1 What You Must Know About Fit\_Downscaler.m

This opening section describes what a user must know before running the Fit\_Downscaler.m function:

1. Please read and respond to the prompt dialog boxes carefully. Each dialog box that appears is important and must be entered correctly for the Fit\_Downscaler.m function to work properly.
2. The amount of memory that is required to run the function depends on the size of your data set. As a rule of thumb, you should have about 8–10 GB of memory available, but less is okay (but slower) if your CMAQ input is 1 GB or smaller.
3. The station measurement data set and the numerical model data set must each contain at least 4 columns with the following information: date, latitude, longitude, and measurement (but not necessarily in that order). The data files may contain more columns of data (e.g. a column of station numbers) but these columns will be ignored in the model fitting as they are not pertinent information.
4. The numerical model data set must NOT have any missing values. For example, if the numerical model was run for 30 days at 3000 centroids, then the numerical model data set must have  $30 \times 3000$  data rows. Otherwise, the Fit\_Downscaler.m function will return an error.
5. The region of the numerical model output must contain the region of monitoring stations. More precisely, if  $\mathcal{D}_C$  is the region in space where the numerical model output is given and  $\mathcal{D}_M$  is the region of space where the monitoring stations are given then it must be the case that  $\mathcal{D}_M \subset \mathcal{D}_C$ . For example, if CMAQ values are given for the eastern part of the United States, the monitoring stations must also be in the eastern part of the U.S.
6. As in #5, prediction can only be done to locations within the numerical model grid. In other words, if  $\mathcal{D}_C$  is defined as in #5 above and  $\mathcal{D}_P$  is the region containing the kriging locations then  $\mathcal{D}_P \subset \mathcal{D}_C$ . If a prediction site, say  $s^*$  is outside  $\mathcal{D}_C$ , then Fit\_Downscaler.m will return a NaN.
7. Negative values for both the numerical model output and monitoring station locations are treated as missing. See #4 in this regard.

## 2 Preliminaries

This section gives an overview of the Fit\_Downscaler.m function. Subsequent sections provide more in depth descriptions of the various components of the function.

## 2.1 Overview

Fit\_Downscaler.m is a MATLAB function that fits a smoothed downscaler with spatially varying weights (similar to that of Berrocal et al. (2010; 2011)) to monitoring station and numerical model data. The code was written on the following platform:

---

MATLAB Version 7.11.0.584 (R2010b)

MATLAB License Number: 359028

Operating System: Linux 2.6.32-131.21.1.el6.x86\_64

Java VM Version: Java 1.6.0\_17-b04 with Sun Microsystems Inc.

Java HotSpot(TM) 64-Bit Server VM mixed mode

---

MATLAB	Version 7.11	(R2010b)
Simulink	Version 7.6	(R2010b)
Aerospace Blockset	Version 3.6	(R2010b)
Aerospace Toolbox	Version 2.6	(R2010b)
Bioinformatics Toolbox	Version 3.6	(R2010b)
Communications Blockset	Version 5.0	(R2010b)
Communications Toolbox	Version 4.6	(R2010b)
Control System Toolbox	Version 9.0	(R2010b)
Curve Fitting Toolbox	Version 3.0	(R2010b)
Database Toolbox	Version 3.8	(R2010b)
Econometrics Toolbox	Version 1.4	(R2010b)
Filter Design Toolbox	Version 4.7.1	(R2010b)
Financial Derivatives Toolbox	Version 5.6	(R2010b)
Financial Toolbox	Version 3.8	(R2010b)
Fixed-Point Toolbox	Version 3.2	(R2010b)
Fuzzy Logic Toolbox	Version 2.2.12	(R2010b)
Global Optimization Toolbox	Version 3.1	(R2010b)
Image Acquisition Toolbox	Version 4.0	(R2010b)
Image Processing Toolbox	Version 7.1	(R2010b)
Instrument Control Toolbox	Version 2.11	(R2010b)
MATLAB Compiler	Version 4.14	(R2010b)
Mapping Toolbox	Version 3.2	(R2010b)
Model Predictive Control Toolbox	Version 3.2.1	(R2010b)
Neural Network Toolbox	Version 7.0	(R2010b)
Optimization Toolbox	Version 5.1	(R2010b)
Parallel Computing Toolbox	Version 5.0	(R2010b)
Parallel Computing Toolbox	Version 5.0	(R2010b)
Partial Differential Equation Toolbox	Version 1.0.17	(R2010b)
Real-Time Workshop	Version 7.6	(R2010b)
Robust Control Toolbox	Version 3.5	(R2010b)
Signal Processing Blockset	Version 7.1	(R2010b)
Signal Processing Toolbox	Version 6.14	(R2010b)
SimBiology	Version 3.3	(R2010b)

SimMechanics	Version 3.2.1	(R2010b)
Simscape	Version 3.4	(R2010b)
Simulink 3D Animation	Version 5.2	(R2010b)
Simulink Control Design	Version 3.2	(R2010b)
Simulink Fixed Point	Version 6.4	(R2010b)
Simulink Verification and Validation	Version 3.0	(R2010b)
Stateflow	Version 7.6	(R2010b)
Stateflow Coder	Version 7.6	(R2010b)
Statistics Toolbox	Version 7.4	(R2010b)
Symbolic Math Toolbox	Version 5.5	(R2010b)
System Identification Toolbox	Version 7.4.1	(R2010b)
Wavelet Toolbox	Version 4.6	(R2010b)

Not all of the above toolboxes are required to run the function. **ONLY THE STATISTICS TOOLBOX IS REQUIRED.**

First, to install the function simply move the **Downscaler** folder which contains the `Fit_Downscaler.m` function and the **Utilities** folder to any location of your choosing. After opening MATLAB, change the current directory to the location of the `Fit_Downscaler.m` function by typing:

```
>> cd TheDirectory/Downscaler
```

where *TheDirectory* is the path name where you saved the **Downscaler** folder and its contents. To invoke the function type:

```
>> Fit_Downscaler
```

at the command prompt. You will then be prompted via dialog boxes for the required information to run the function (see Section 3). Be prepared with the following information:

1. Locations for the monitoring station and numerical model data files.
2. Desired MCMC settings (e.g. number of draws, burn, and thinning).
3. Path name (folder) where output is going to be written.
4. Data formats for the monitoring station and numerical model data files. For example, delimiter, number of headerlines, column name and data type (e.g. numeric or string).

`Fit_Downscaler.m` will write to the specified output directory the following five subdirectories: **MCMC**, **RESULTS**, and **VALIDATION**. The **MCMC** folder will contain successive MCMC draws of the parameters  $\beta_0$ ,  $\beta_1$ ,  $\sigma_{\beta_0}^2$ , and  $\tau_Y^2$  which can be used to assess convergence of the MCMC sampler (see Section 2.2). The **RESULTS** folder contains the mean and standard deviation of the posterior predictive distribution for the predicted locations if the prediction option is chosen (see Section 4.2 for more information). Finally, the **VALIDATION** folder contains the coverage, square error, and bias from the validation sample if the validation option is chosen (see Section 4.1 for more information). See Section 5 for specific information on the output from the `Fit_Downscaler.m` function.

## 2.2 Methodology

Let  $Y(\mathbf{s}, t)$  be the observed monitoring station data at location  $\mathbf{s}$  on day  $t$ . Furthermore, let  $X(B_k, t)$  be the observed numerical model output for grid block  $B_k$  on day  $t$ . The smoothed downscaler with spatially varying weights (hereafter referred to as the SVW model) assumes,

$$Y(\mathbf{s}, t) = \beta_{0,t} + \beta_0(\mathbf{s}, t) + \beta_{1,t}\tilde{X}(\mathbf{s}, t) + \epsilon(\mathbf{s}, t), \quad (1)$$

where  $\beta_0(\mathbf{s}, t)$  is a mean zero Gaussian process with covariance function,

$$\begin{aligned} \text{Cov}(\beta_0(\mathbf{s}, t), \beta_0(\mathbf{s} + \mathbf{h}, t + u)) &= \sigma_{\beta_{0,t}}^2 \exp\{-\phi_{\beta_{0,t}}\|\mathbf{h}\|\} \mathbb{I}_{\{u=0\}} \\ &= \mathbb{C}_{\beta_{0,t}}(\mathbf{h}, u \mid \phi_{\beta_{0,t}}), \end{aligned} \quad (2)$$

and  $\epsilon(\mathbf{s}, t)$  is a white noise Gaussian process with variance  $\tau_t^2$ . Intuitively, the SVW model assumes independence across time while the spatial dependence is governed by a Gaussian process with exponential covariance function. The predictors  $\tilde{X}(\mathbf{s}, t)$  in (1) are defined as,

$$\tilde{X}(\mathbf{s}, t) = \sum_{k=1}^K w_k(\mathbf{s}, t) X(B_k, t), \quad (3)$$

where,

$$w_k(\mathbf{s}, t) = \frac{\exp\{-\phi_w\|\mathbf{s} - \mathbf{c}_k\|\} \exp\{Q(\mathbf{c}_k, t)\}}{\sum_{k=1}^K \exp\{-\phi_w\|\mathbf{s} - \mathbf{c}_k\|\} \exp\{Q(\mathbf{c}_k, t)\}}, \quad (4)$$

$K$  is the total number of numerical model grid cells,  $\phi_w$  is a decay parameter,  $\mathbf{c}_k$  is the centroid for the  $k^{th}$  grid cell, and  $Q(\mathbf{s}, t)$  is a mean zero Gaussian process with covariance function  $\mathbb{C}_Q(\mathbf{h}, u \mid \phi_Q)$  which is defined similarly to (2). In this way  $\exp\{Q(\mathbf{c}_k, t)\}$  defines the multiplicative increase to the spatial weight given by  $\exp\{-\phi_w\|\mathbf{s} - \mathbf{c}_k\|\}$ .

The model also allows transformations (log or square root) of the monitoring and CMAQ data to be used in the model. In terms of the notation above, let  $Y^*(s, t)$  and  $X^*(B_k, t)$  be the observed monitoring and CMAQ data. We then use transformations  $g_1$  and  $g_2$  (typically the same transformation) to get  $Y(s, t) = g_1(Y^*(s, t))$  and  $X(B_k, t) = g_2(X^*(B_k, t))$ .  $Y(s, t)$  and  $X(B_k, t)$  are then used as described above. Prediction and validation still occurs on the original scale of the data  $Y^*(s, t)$ .

## 2.3 Fitting the Model

This section describes the various parameters of the model and how estimation of these parameters are handled in Fit\_Downscaler.m. Markov chain Monte Carlo methods are used to obtain a sample from the posterior distribution of all model parameters. Draws are obtained using a Gibbs sampler where the Metropolis-Hastings accept-reject algorithm is used where necessary. See below for more details.

1.  $(\tau_t^2, \beta_{0,t}, \beta_{1,t})$ . Vague conjugate prior distributions are assumed for these parameters. Additionally,  $\{\beta_{0,t}\}$  and  $\{\beta_{1,t}\}$  are assumed to be independent in time. Thus, conditional on  $\{\tilde{X}(\mathbf{s}, t)\}$ ,  $(\tau_t^2, \beta_{0,t}, \beta_{1,t})$  can be sampled via composition by first sampling  $\tau_t^2$  then sampling  $(\beta_{0,t}, \beta_{1,t})$  conditional on the obtained value of  $\tau_t^2$ .

2.  $\{\beta_0(\mathbf{s}, t)\}$ . The prior distribution for  $\beta_0(\mathbf{s}, t)$ , as stated above, is a mean zero Gaussian process with exponential covariance function and decay parameter  $\phi_{\beta_0, t}$ . Conditional on all the parameters, the entire vector of  $\{\beta_0(\mathbf{s}, t)\}$  can be sampled from its Gaussian complete conditional distribution.
3.  $\sigma_{\beta_0, t}^2$ . Using the conjugate inverse gamma distribution,  $\sigma_{\beta_0, t}^2$  can be drawn directly from its complete conditional distribution.
4.  $\phi_{\beta_0, t}$ . This parameter represents the rate of decay of the spatial correlation in  $\{\beta_0(\mathbf{s}, t)\}$ . Specifically, as  $\phi_{\beta_0, t}$  increases, the spatial correlation of  $\beta_0(\mathbf{s}, t)$  decreases. As discussed in Zhang (2004), decay parameters for spatial process are particularly difficult to estimate. As such, a grid search for the spatial range with discrete prior which places mass at (10, 20, ..., 90) percent of the maximum observed distance between  $Y(\mathbf{s}, t)$  and  $Y(\mathbf{s}', t)$  is used. After performing the grid search,  $\phi_{\beta_0, t}$  is fixed at the most likely value (from the grid search) for the remainder of the MCMC algorithm.
5.  $(\phi_w, \phi_Q)$ . These parameters represent the decay parameters for the weights  $w(\mathbf{s}, t)$  used to weight various neighboring numerical model grid cell observations. Berrocal et al. (2011) note that very little information about these parameters is available. As such,  $\phi_w$  and  $\phi_Q$  are fixed such that the weights (correlation) are essentially zero beyond a distance of 3 numerical model grid cells. In testing the `Fit_Downscaler.m` function, these values seem to work well.
6.  $\{Q(\mathbf{c}_k, t)\}$ . Each  $Q(\mathbf{c}_k, t)$  defines a multiplicative increase in the weight assigned to the numerical model value  $X(B_k, t)$ . Intuitively, if  $Q(\mathbf{c}_k, t)$  is large then  $X(B_k, t)$  will have a large effect on  $Y(\mathbf{s}, t)$ . Notice that there is one  $Q(\mathbf{c}_k, t)$  for each numerical model (i.e. CMAQ) centroid. This represents, potentially, a very large number of  $Q(\mathbf{c}_k, t)$ . As such, predictive processes (see Finley et al. 2009) are used to reduce the dimensionality. Specifically,  $Q(\mathbf{c}_k, t) \equiv \mathbb{E}(Q(\mathbf{c}_k, t) \mid Q(\check{\mathbf{c}}_j, t))$  where  $\{Q(\check{\mathbf{c}}_j, t)\}$  are a set of sparsely chosen locations with the number of  $Q(\check{\mathbf{c}}_j, t)$  being substantially less than the number of  $Q(\mathbf{c}_k, t)$ . In this way, all  $Q(\mathbf{c}_k, t)$  can be updated by updating relatively few  $Q(\check{\mathbf{c}}_j, t)$ . The `Fit_Downscaler.m` function uses 100  $Q(\check{\mathbf{c}}_j, t)$  located on a grid from the lowest (in terms of latitude and longitude) to the highest observed grid cell. To update  $\{Q(\check{\mathbf{c}}_j, t)\}$ , a Metropolis-Hastings step is used with proposal distribution equal to the prior distribution.

## 2.4 Formatting Your Data

Both the numerical model data and the monitoring station data must contain at least 4 columns - latitude, longitude, date, and value. These 4 columns can be arranged in any order as the function will prompt the user for the data format (see Section 3 for more details). **The data files must be either comma, space, or tab delimited and have extensions .txt or .csv.**

*Latitude.* The latitude column contains the latitude locations (in degrees) of the monitoring station or numerical model centroid. This column must contain only double precision

numbers. “Missing” values such as “NA” or “NaN” are not allowed and will result in an error.

*Longitude.* The longitude column contains the longitude locations (in degrees) of the monitoring station or numerical model centroid. This column must contain only double precision numbers. “Missing” values such as “NA” or “NaN” are not allowed and will result in an error.

*Date.* The date column is a string specifying the date the measurement was taken. The date column must be in one of the formats mentioned in Section 3.5.

*Measurement.* The measurement column is a column of double precision numbers. Values less than zero are treated as missing and will be discarded from the analysis. The numerical model data measurement should have all positive numbers.

As mentioned in Section 1, the region of numerical model output must contain the region of monitoring stations. In other words, for each monitoring station, it is assumed that the numerical model has been run in the same region such that each monitoring station location is contained within a numerical model centroid included in the data files.

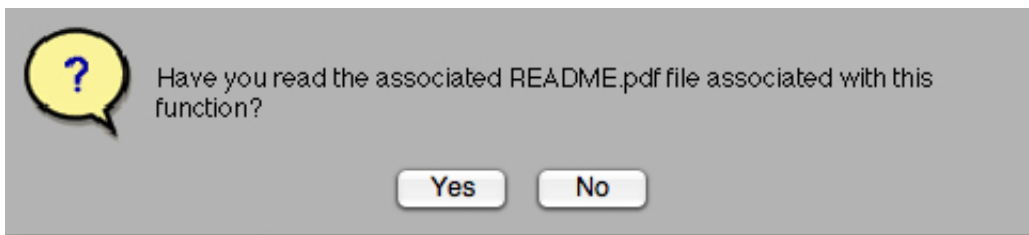
## 3 Initial Call to Fit\_Downscaler.m - Dialog Boxes

After typing,

```
>> Fit_Downscaler
```

at the MATLAB command prompt, a series of dialog boxes will appear. This section describes those dialog boxes in more detail.

### 3.1 README Dialog Box



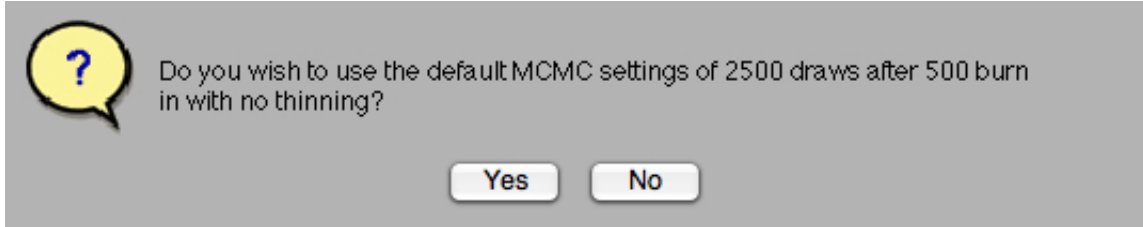
The first box to appear is the README dialog box (above). This box simply prompts the user to make sure they have first read this manual before running the function. If not, the function will not execute.

### 3.2 File Browser

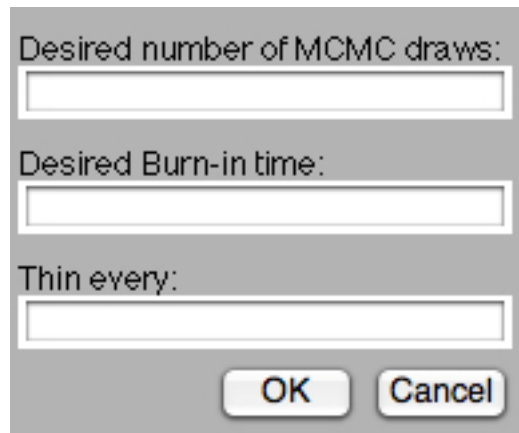
If “Yes” is selected in the README Dialog box, two browser windows will appear (the browser windows will look different depending on the platform (e.g. windows, mac, unix) you are using and, as such, are not pictured here). The first is titled “Select the monitoring station data file” wherein the user will browse and select the file that contains their monitoring

station data. The second is titled “Select the numerical model data file” wherein the user will browse and select the file that contains their numerical model data.

### 3.3 MCMC Settings Dialog Box

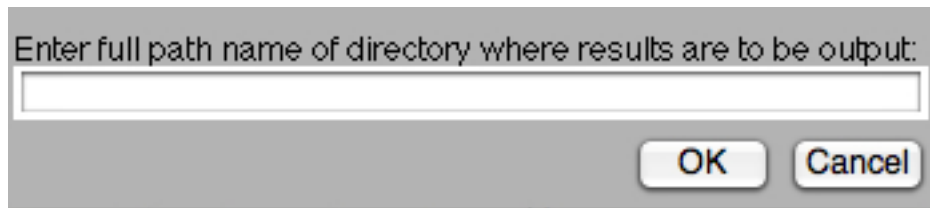


After selecting the data files in the browser windows, the MCMC setting dialog box above will appear. The default setting is to obtain 2500 draws from the posterior distribution after an initial burn-in period of 500 draws and no thinning (i.e. keep every draw). If the user selects “no” then the following dialog box will appear:



In each of the provided spaces, the user selects the number of draws to obtain from the posterior after a specified burn-in time. Optionally, the user can specify a thinning value which will keep every  $t^{th}$  draw where  $t$  is the thinning value. For example, if  $t = 2$  then every  $2^{nd}$  draw will be kept.  $Thin = 1$  corresponds to no thinning. **We strongly recommend that the default of 2500 draws and 500 burn be minimum values.**

### 3.4 Output Dialog Box



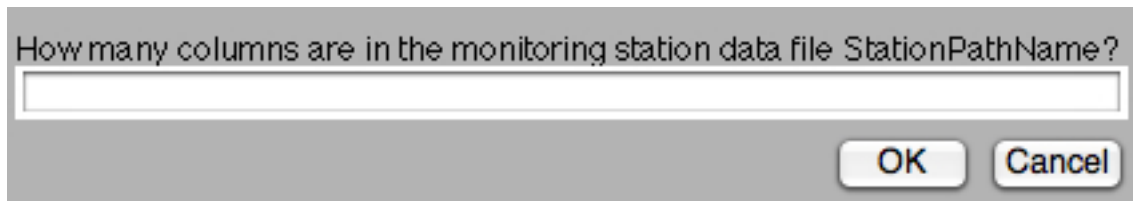
Following the MCMC dialog boxes, this dialog box prompts the user to enter a valid folder where the output of the function is to be written. If the specified directory does not exist,



the user will be prompted again to enter a valid output folder. Once a valid output folder has been supplied, various folders will be created and output from the function will be written to the created folders (see Section 5 for details on the output of the function).

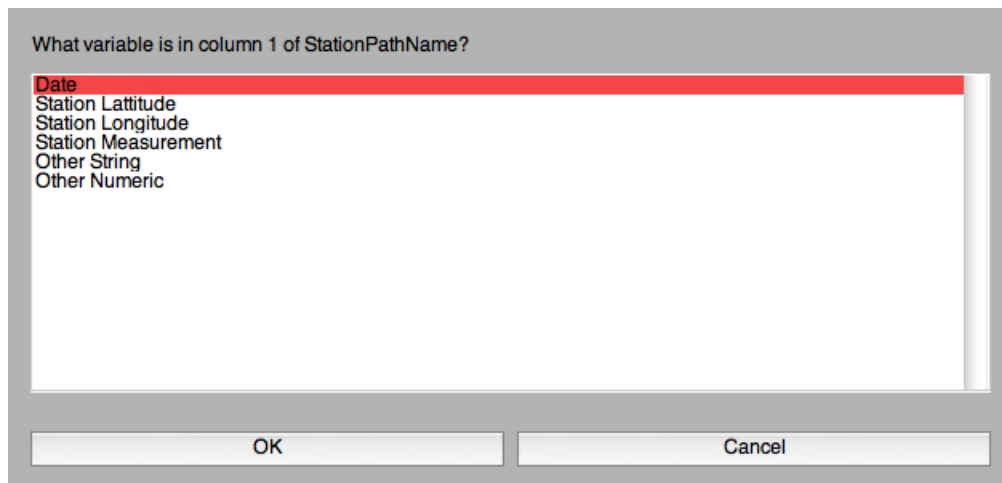
### 3.5 Data Format Dialog Boxes

As this point, a series of dialog boxes will appear which prompt the user to supply the necessary information to read in the data. The first box to appear is as follows:



In this box, the user should enter the **total number of columns** in the file specified in the pathname given to the function in Section 3.2. If the user supplies an incorrect number of columns, the function could draw an error.

After supplying the number of columns, the program will prompt the user for the content in each column by looping through the following dialog box:



The `Fit_Downscaler.m` function assumes that the date column of the data files is a string. Station latitude, longitude, and measurement should all be double precision and contain no strings or letters. If there are more than 4 columns in the specified file (e.g. station number) then use the “other string” or “other numeric” fields to tell the program how to handle these columns. For example, if column 5 of your data set contains the state where the monitoring station is located then choose “Other String” for column 5. If column 6 contains a station number, then choose “Other Numeric” for column 6.

**NOTE:** The data files specified in Section 3.2 above must both contain at least 4 columns with the following information: date, latitude, longitude, and measurement (but not necessarily in that order). If one of these fields is not available then the function will return an error.

After looping through and prompting the data type for each column in the path names specified in Section 3.2, the user will be prompted for the number of header lines by the following dialog box:

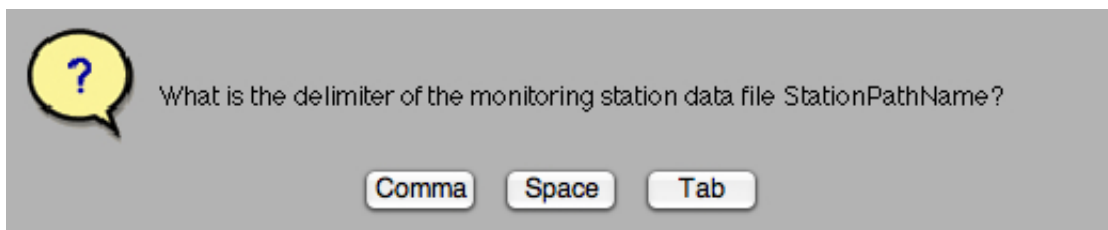


How many header lines are in the monitoring station data file StationPathName

OK Cancel

The number of header lines is equivalent to the number of lines that will be skipped when reading in the data.

Next, the user will be prompted to specify the delimiter for the data files:

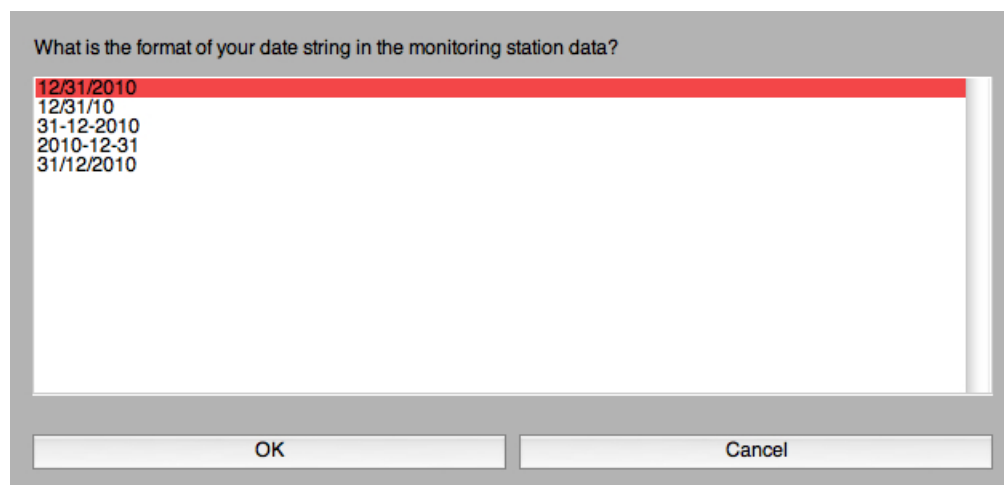


What is the delimiter of the monitoring station data file StationPathName?

Comma Space Tab

The files given in Section 3.2 can be delimited by comma, space, or tab. Any other delimiter is not allowed and will result in an error.

Finally, the user is prompted for the format for the date string used in the data file with the following dialog box:



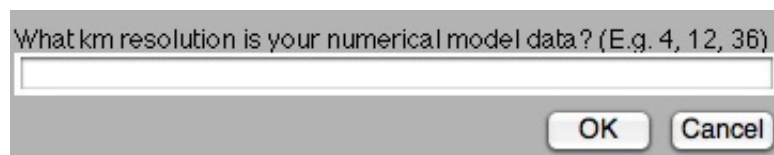
What is the format of your date string in the monitoring station data?

- 12/31/2010
- 12/31/10
- 31-12-2010
- 2010-12-31
- 31/12/2010

OK Cancel

If the date format is not contained in this list, please reformat the date string to match.

The same set of dialog boxes will appear to prompt the user for the format of the numerical model data file. Following the above prompts for the numerical model data file, the user will be prompted for the resolution of the numerical model data by the following dialog box:



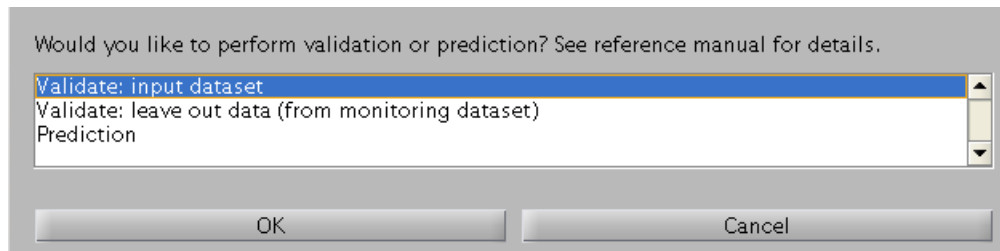
What km resolution is your numerical model data? (E.g. 4, 12, 36)

OK Cancel

This will determine the values of  $\phi_w$  and  $\phi_Q$  above.

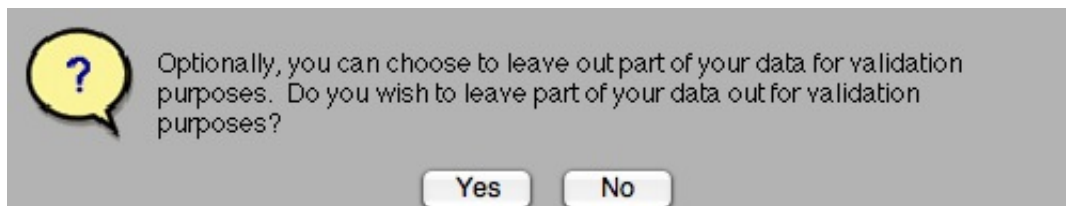
### 3.6 Options Dialog Boxes

At this point, the user can select from two option boxes: validation and kriging locations. The box is given by:



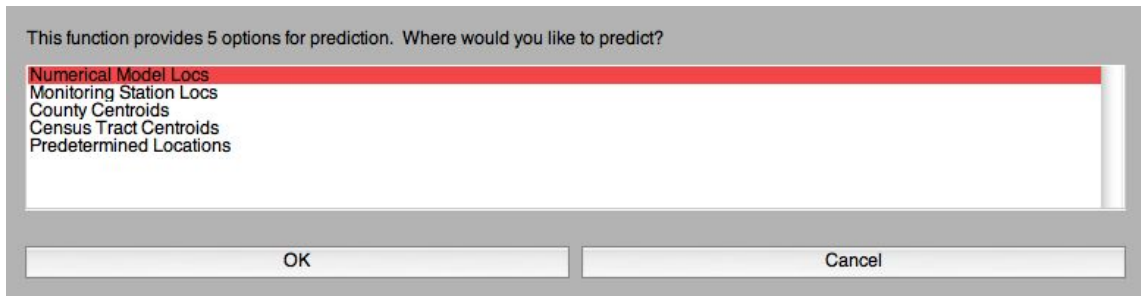
and prompts the user to choose either validation or prediction. In both validation choices, various validation statistics (coverage, bias, root mean square error) will be calculated (see Section 4.1) and output. The first option allows the input of a second set of monitoring data to be used for validation. This validation data should be exclusive from the first monitoring data input (no data for same location on same day). The user will be prompted to choose whether the data format is exactly the same as the monitoring data (e.g., number and content of columns, date format, etc.) or not. If the format is different, then the user will be prompted with dialog boxes as before.

The second validation option allows a portion of the data wants to be left out for validation purposes. This option prompts the user with the following

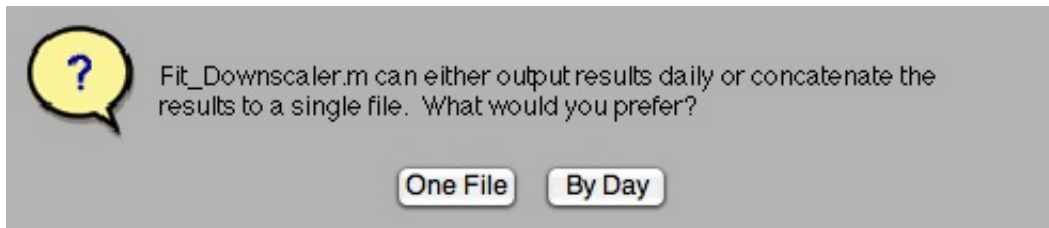


and allows the user to input what percentage of the data should be held out for calculation (e.g., input 10 for ten percent validation).

If prediction is chosen, then validation will not be performed and the user will be able to choose a set of prediction options. Under the prediction option, the main output from `Fit_Downscaler.m` is a list of predictions and uncertainty measurements for a set of prediction locations. The user has five options when it comes to predictions. The user can obtain a prediction and measure of uncertainty for (i) all numerical model centroids, (ii) all monitoring station locations, (iii) centroids of all counties in the U.S., (iv) centroids of all census tracts in the U.S., or (v) a user specified set of prediction locations. At this point, the user will be prompted for which locations a prediction and measure of uncertainty is desired via the following dialog box.



If “Predetermined Locations” is selected, the user will be prompted via dialog boxes similar to those in Section 3.5 for the file path of a file containing the latitudes and longitudes of the desired prediction locations. Output can be written to a single file or to multiple files containing the predictions by day (if multiple days are included). The user is prompted for this via the following dialog box:



It is highly recommended that results are written “By Day” because “One File” may be a very large file that is hard to manipulate. Outputting results “By Day” will contain a lot of files but the files will be smaller.

## 4 Options for Fit\_Downscaler.m

The Fit\_Downscaler.m function prompts for choosing validation or prediction. The user can perform validation by inputting a validation dataset or choosing to leave out part of the data and use it for validation purposes. Second, the user can supply a list of latitudes/longitudes (in degrees) for which the code will produce a predicted value and an associated measure of uncertainty. This section provides details on each option.

### 4.1 Validation

If the user chooses to input a validation dataset, the model will predict the data value for each observation in the validation dataset. If the user elects to leave out part of the data for validation purposes, the user will be prompted for a percentage of daily data to leave out. For each day of data, the model will leave out the desired percentage of data. In both cases, the model will output (both on-screen and in the VALIDATION output folder) bias, squared error, and coverage (calculated on the original scale of the data if a transformation was used). The VALIDATION folder will contain a single file called ValidationResults.csv which contains the following 6 columns:

1. *Date*. Validation date (string).

2. *Latitude*. Validation location latitude.
3. *Longitude*. Validation location longitude.
4. *Bias*. The fourth column gives  $\hat{y} - y$  where  $\hat{y}$  is the mean of the posterior predictive distribution and  $y$  is the observed value. The total bias over all days can be obtained by averaging the fourth column.
5. *Squared Error*. The fifth column contains the squared error for that site. In other words, the third column gives  $(\hat{y} - y)^2$ . The total mean square error can be obtained by taking the mean of the fifth column.
6. *Coverage*. The sixth column contains 1 or 0 indicating if a 95% prediction interval contained the observed value. The total coverage for the left out data can be obtained by averaging the sixth column.

## 4.2 Prediction

If the user chooses to supply a list of predetermined latitude-longitudes (in degrees) to which the model will provide a predicted estimate and a measure of uncertainty, this list must be contained within the numerical model output. For example, `Fit_Downscaler.m` assumes that each location in the supplied list is contained within a numerical model centroid within the file provided in Section 3.2 above. If not, the function will output a NaN value for that predicted location.

If the user elected to output to a single file, the RESULTS folder will contain a single file called Predictions.csv which contains the following columns:

1. *Date*. Prediction date (string). This will be output ONLY if “One File” is written to results folder (see Section 3.6).
2. *Federal Information Processing Standard (FIPS) code*. 11-digit FIPS code ([http://quickfacts.census.gov/qfd/meta/long\\_fips.htm](http://quickfacts.census.gov/qfd/meta/long_fips.htm)) that denotes state FIPS code (digits 1–2), county FIPS code (digits 3–5), census tract base (digits 6–9), and census tract suffix (digits 10–11). (this is only written if “Census Tract Centroids” was chosen for prediction locations).
3. *County State*. Name of state in which county resides (this is only written if “County Centroids” was chosen for prediction locations).
4. *County Name*. Name of county (this is only written if “County Centroids” was chosen for prediction locations).
5. *Latitude*. Prediction location latitude.
6. *Longitude*. Prediction location longitude.
7. *Prediction*. Point prediction (on the original scale, if a transformation was used).
8. *Uncertainty*. The posterior standard deviation (on original scale). An approximate 95% credible interval prediction  $\pm 2 \times (\text{SD})$ .

## 5 Output from a call to Fit\_Downscaler.m

The Fit\_Downscaler.m function will create the following 3 subdirectories within the output folder specified in the output dialog box (see Section 3.4 of this manual):

1. MCMC. This folder contains successive MCMC draws for the global parameters  $\beta_{0,1}$ ,  $\beta_{1,t}$ ,  $\sigma_{\beta_0}^2$ , and  $\tau_t^2$  for each day of the analysis (in that order). These draws should be plotted to assess convergence of the MCMC algorithm.
2. RESULTS. If prediction is chosen, this folder contains the predicted values for the defined prediction locations (see Section 3.6) along with the standard deviation of the posterior predictive distribution. The standard deviation of the posterior predictive distribution is a measure of uncertainty. An approximate 95% predictive interval for the predicted value can be computed as value  $\pm 2*(SD)$ . If a transformation of the input datasets is used, the predicted values and uncertainties are on the original scale. If output is to be written to a single file, RESULTS will contain a single file (`Predictions.csv`) otherwise it will contain daily files. See Section 4.2 for more details on the column names of the files in the RESULTS folder. If validation is chosen, this folder will remain empty.
3. VALIDATION. If validation is chosen, this folder contains the results from the validation analysis. See Section 4.1 for more details. If prediction is chose, this folder will remain empty.

## 6 Troubleshooting

Please direct all questions to David Holland (Holland.David@epamail.epa.gov).

**Disclaimer:** The United States Environmental Protection Agency through its Office of Research and Development funded and managed the research described here. It has been subjected to Agency review and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation by EPA for use.

## REFERENCES

- Berrocal, V. J., Gelfand, A. E., and Holland, D. M. (2010). A spatio-temporal downscaler for output from numerical models. *Journal of Agricultural, Biological, and Environmental Statistics*, 15:176–197.
- Berrocal, V. J., Gelfand, A. E., and Holland, D. M. (2011). Space-time data fusion under error in computer model output: an application to modeling air quality. Accepted in *Biometrics*, online version: <http://onlinelibrary.wiley.com/doi/10.1111/j.1541-0420.2011.01725.x/abstract>.

- Finley, A. O., Sang, H., Banerjee, S., and Gelfand, A. E. (2009). Improving the performance of predictive process modeling for large datasets. *Computational Statistics and Data Analysis*, 53:2873–2884.
- Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261.