Keith Arora-Williams
Landscape Hydrology Final Project
August 15, 2016

Introduction

The overall goal of my research is to try to bring together new information about the distribution of microbial genes & genomes in aquatic environments with well-understood chemical species dynamics. Many of the best biogeochemical models developed for the purpose of simulating chemical reactions and transport processes in water bodies were formulated in the pre-genomics era. As this new information becomes more available and comprehensible, it is the task of those on the forefront to reconcile the presence/absence of enzyme-coding nucleic acid sequences (or the microbes that encase them) with the chemical transformations assigned to them by biogeochemical models, throughout the body of the water column. By doing so, it is possible to get a better understanding of the stability of microbial life in a dynamic environment that is host to storm water runoff and sewage overflow deposition, as well as human recreation and migrations of various species of aquatic life.

Prehiem et al. (2016) was implemented a biogeochemical model inspired by BIORXNTRN (Hunter et al. 1998) to simulate a specific set of primary & secondary microbial metabolic redox reactions based on initial values measured at 1 meter depth intervals for Upper Mystic Lake in Winchester, MA. The model uses theoretical energetic yields based on redox potentials to determine the order in which nutrients become depleted at each depth layer. Primary oxidation rates follow a formulation informed by the relative favorability of electron acceptors. Secondary oxidation rates follow simple mass action rate forms. Transport between meter-high compartments is achieved in this simulation by two processes: (1) All chemical species are allowed to diffuse into adjacent compartments by a Fickian diffusion type relationship. (2) biomass-associated carbon and oxidized iron settle downward at a fixed, equivalent rate. The outside world is modeled by constant source terms: oxygen and biomass are added in the uppermost compartment (at the thermocline), while methane is added in the lowermost compartment (at the sediment). The resulting set of ordinary differential equations is solved numerically. Transport is modeled compartment-by-compartment, that is, using ordinary differential equations rather than partial differential equations as in BIORXNTRN. Some simulated chemical species which consist of multiple chemical species in nature are simplified to one summary term (e.g., the modeled oxidized sulfur species includes hydrogen sulfide, bisulfide, and sulfide), while others are omitted entirely manganese. Precipitation-dissolution, acid dissolution, and adsorption reactions are also ignored.

To address some of these shortcomings, I sought out existing biogeochemical models built upon sophisticated hydrological models. The requirements for such a model are (1) it simulates all the reactions programmed into the current Mystic Lake Model (2) the reactions simulated are customizable and modular (3) it is actively supported and widely used by the lake modeling community, as evidenced by current literature (3) it simulates transport by advection, convection, settling, and optionally molecular/turbulent diffusion. I selected the interoperable Aquatic Eco-Dynamics & General Lake Models as they satisfied all these requirements (Hipsey, & Bruce, 2013)

In the following paragraphs, I will summarize my efforts to simulate the observed temperature distribution of the water column in Mystic Lake across 10 time points in from August 2012 to May 2013 using the General Lake Model. The purpose of this specific document is to try to convey a sense of which model inputs, among input data types and model parameters, exert the greatest influence over vertical mixing i.e. diapyncal transport.

Methods

Location

The Mystic Lakes are located at 42.4317° N, 71.1483° W, within the boundaries Winchester, Arlington, and Medford, MA. They both lie 1 meter above sea level. The Aberjona River and associated watershed feeds into Upper Mystic Lake, which flows over the Mystic Dam into Lower Mystic Lake. Lower Mystic Lake drains into the Mystic River, which flows for ~10 miles until it reaches Boston Harbor and the Atlantic Ocean. A map of the location and surrounding landmarks is provided in the Supplementary Information

Time Series Data & Sources

Various forms of meteorological and water-quality data is required to run the GLM. Daily summaries for wind speed, air temperature (min/max), precipitation, & snowfall data recorded from a land-based station at Boston Logan Airport was sourced from the NOAA National Centers for Environmental Information site (*http://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00014739/detail*).

Longwave & Shortwave radiation, relative humidity, cloud cover fraction, and skin-surface temperature data were obtained from the NASA Clouds and the Earth's Radiant Energy System archive (*http://ceres.larc.nasa.gov/order_data.php*). A bounding box measuring 2 degrees in length and width, centered at Upper Mystic Lake was used to download the data. Once downloaded, measurements from the single closest geoposition were retained. This process is recommended as each data type is indexed at different spatial scales and a restrictive bound-box can yield irregularly fragmented data.

Column-Averaged Humidity data was taken from the Level 2 "SSF" data product and the other data types were taken from the Level 3 "SYN Deg" data product. The specific type of radiation data used was the daily surface computed shortwave/longwave downward radiation. The cloud cover fraction data was compared between CERES data products and with a comparable data product distributed by the Goddard Earth Sciences Data and Information Services Center (*http://daac.gsfc.nasa.gov/*). The Cloud Cover data in the Level 3 product more closely resembled the GESDISC data (R=0.73), so it was utilized.

The discharge, salinity, & water temperature data was obtained from the USGS Water Data archive. A continuous record of discharge was available for the Aberjona River dating back from 1937 up to the present (*http://waterdata.usgs.gov/nwis/inventory?agency_code=USGS&site_no=01102500*). A daily record of salinity and water temperature is not available for the Aberjona River, however data from nearby Hobbs Brook was substituted. The water & air temperature values taken from Hobbs Brook and one additional nearby water body were plotted and the comparison revealed minimal differences (Supplementary Info).

A summary of the processed data that was ultimately used in model simulations is shown in Table 1.

Table 1: General Lake Model Input Data Summary: Radiation, air temperature, relative humidity, cloud cover, and water temperature were provided in the correct units. The multiplier used to convert discharge data from cubic feet per second to Megaliters per day was 2.44658. Both forms of precipitation were scaled up from tenths of millimeters to meters per day. mg/L was recommended as the unit for salinity data, however no clear conversion method was available, so the units in which the data was provided by the USGS were used.

Data Set	**Mean**	**Std**	**(Max, Min)**	**Zeros**	**Date-Range**	**Units**
ShortWave	137.96	80.50	(312.88, 6.80)	0	2005-08-01 2014-06-30	W/m^2
LongWave	307.90	56.90	(423.13, 169.37)	0	2005-08-01 2014-06-30	W/m^2
Air Temperature	11.16	9.53	(33.3, -16.95)	40	2000-01-01 2015-01-01	C

Data Set	Mean	Std	(Max, Min)	Zeros	Date-Range	Units
Relative Humidity	38.68	12.78	(82.53,10.64)	0	2010-01-01 2014-01-01	%
Cloud Cover	17.02	20.65	(99.0, 0.0)	315	2010-01-01 2014-01-01	Fraction
Wind Speed	4.87	1.67	(13.9, 1.0)	0	2000-01-01 2015-01-01	m/s
Rain	0.0030	0.008	(0.109, 0.0)	3570	2000-01-01 2015-01-01	m/day
Snow	0.0033	0.023	(0.599, 0.0)	5150	2000-01-01 2015-01-01	m/day
Inflow Discharge	31.73	47.43	(1420.0, 0.25)	0	1939-04-17 2016-03-31	ML/day*
Inflow Water Temp.	12.12	7.34	(26.4, 0.2)	0	2005-10-02 2016-06-29	C
Inflow Salinity	763.22	169.37	(1760.0, 414.0)	0	2005-10-02 2016-06-29	uS/cm **

Morphometry

All absolute morphometric data was obtained using tools presented in the My Maps interface of Google Maps. The basin length and width were measured to be 1012 and 536 meters, respectively. The morphometry is set in the model based on a map of the depth-area contours at 10 ft intervals (Supplementary Info). The relative areas between each contour was calculated based on a pixel counts from binarized images. These relative areas were converted to absolute areas based on the area of lake surface shown in satellite images Google Maps measured with the program ImageJ. A simple estimate of the volume of the lake can be obtained by taking the sum of the 10 ft x the area vector, which yields 8.9e6 cubic meters. This value is consistent with the modeled steady-state Lake volume of 9.76e6 (Figure 1). The slight difference can be attributed to the addition of a conical volume between the lowest area slice and the distance to the lake nadir, which is a parameter provided to the model..

Program Operation

The General Lake Model is configured using four text files. The first three text files contain the meteorological & flow data as comma-separated values files with specific column names defined the the GLM documentation. Data obtained from the USGS/NOAA/NASA in various formats was imported, processed, interpolated, down-sampled as necessary, and written out in properly formatted CSVs using custom scripts written in Python 2.7 and available on GitHub (*https://github.com/karoraw1/GLM_Wrapper.git*).

The last text file required to run the model is a configuration file containing parameters and arrays of initial values named `*glm2.nml*`. Four such files were provided as examples with the model software and all were reviewed. A selection of the simulation parameters, their definitions, and the value selected in the best model runs is provided in Appendix I. To ensure appropriate formatting, one such example file was read into the LakeModel object and compared with the format of each parameter written out into the `*glm2.nml*` file for Upper Mystic Lake to ensure valid comprehension by the GLM executable. This task is performed by the `*LakeModel.import_config*()` function. This helped in observing the placement of whitespace characters like newlines, tabs, and spaces.

Performance testing was completed using the Linux & Mac OS X binaries of the GLM executable. It was revealed that the latter binary was compiled to support recursive FORTRAN computation, which reduced the total run time for each simulation from ~572 seconds to about 1.5 seconds.

The *initial* parameter values were selected based on provided recommendations and site-specific measurements where available. Early testing revealed that a couple of the boolean parameters used to disable all meteorological forcing, as well as those offering support for non-neutral atmospheric stability and deep mixing were non-functional.

To improve upon the *initial* values, a parameter value space was defined. The parameters of the GLM include a mix of boolean, integer, and continuous parameters, so it seemed simplest to specify the possible choices for each explicitly in a text file (shown with explanation in Appendix II). All booleans could be set to true or false. Integer dummy variables could be specified as a list of their possible values. Scalar parameters were defined in terms of their upper limit, lower limit, and the increments of a linear array of of possible values between them. Finally, parameters & initial conditions defined in the configuration file as arrays (e.g. height-area definitions & water column temperature initial values) could be scaled up or down by a coefficient, which used the same format as scalar parameters. By defining 40 of the model's parameters as ranges instead of specific values, a total of ~1100 valid parameter/value pairs were made available to the model for rapid calibration. The total size of the unique model runs possible in this parameter space is somewhere in the neighborhood of 1e+53 to 1e+55.

To run a simulation, the GLM executable is launched in a directory containing the four required text files. The GLM Wrapper program contains class object definitions for each data type and for each lake simulation. When a Lake object is instantiated, a new directory & `*glm2.nml*` file is created automatically. The data files obtained from the archives listed above are placed in either the `weatherData` or `waterData` folders. Data objects can then be instantiated, which will import the data from those folders into memory. Once the required data objects are fed into the Lake object, the main simulation directory is populated with the required CSV files. The `*run_model*`, `*pull_output*`, and `*score_variant*` functions can then be called, in that order.

The `*read_variants*` & `*write_variant_configs*` can be called after an initial run is performed to read the parameter value space definition file. This creates a directory structure within the original Lake object directory. Each new directory contains the variant configuration file and symbolic links that point to the original data files, to minimize the hard-drive footprint. Variant simulation definitions can be run in parallel or in series, which cuts overall run time of thousands of simulations to approximately two sequential simulations, if sufficient RAM is provided.

Calibration

The first attempt at calibration was performed using a path-dependent, grid-based traversal of a decision tree to a local minimum. This method is alternatively called the greedy algorithm in the following descriptions, as it greedily selects the greatest improvement at each step. The algorithm proceeds as follows: (1) simulation is performed with the *initial* values (2) the simulation is scored by calculating the Nash Sutcliffe Efficiency of the predicted temperatures distribution across time & depth with measured values (3) Each of the ~1100 possible alternative parameter values is tried, while all the other parameters remain at their initial setting, and the NSE calculated. (4) A decision is made to modify a single parameter to the value that produced maximum relative improvement in the NSE among all ~1100 simulations (5) Steps (3) & (4) are then repeated until the relative improvements by parameter modification drops below 0.1%.

Upon consideration, two major shortcomings of this approach exist. The first is that there is no good reason behind modifying the parameter value with the largest reduction in error. The reason that this approach was selected at all was that it yielded results in the shortest amount of time. The degree of improvement appeared to fall dramatically in successive rounds of the algorithm. A closed-form solution that could improve this approach would be (1) select multiple parameter-value pairs that yield a reduction in observed error for modification, based on a threshold of the ratio of the input & output variances, i.e. the variance of the observed changes in error over the variance of the particular parameter's value space (2) measure the potential reduction in error from all possible parameter modifications after the initial modification (3) eliminate all but one of the initial modifications, only keeping those that exhibit the greatest cumulative improvement across both modifications. The increase in computational steps required for this approach is proportional to the of multiple parameter-values pairs to explore in the (2) step. The number of multiple parameter-value pairs explored in the second step could be modified to match the expected complexity of the multiple-parameter interactions. An illustrative example of how this approach could yield superior results would be the case where two interrelated parameters are set to values that cause the simulation to produce high error values. Modifying each parameter in isolation might not make much of a difference to the observed error, however if one of the two is selected for exploration in step (1), the improvement shown by joint modification would be accessible in step (2). This change represents an ideological shift from traversing a decision tree, to exploring a decision forest.

The second shortcoming of this approach is that any effort to calibrate the model is restricted to a path leading away from the initial conditions manually selected by the user. By randomizing parameter selection, a wide variety of starting points would be chosen, which doesn't prevent the possibility of getting caught in a local minimum, but it allows for many different local minima to be explored. The grid-search will predictably get trapped in the closest identical minimum to its starting point. Based on this consideration, the Hornberger-Spear-Young method for sensitivity analysis was performed, as it is a randomized and thus path independent method (Beven, 2011). 17,503 simulations in which 36 parameter values were selected at random from the parameter-value space defined in Appendix II. The highest scoring 5% (n=668) & 1% (n=182) of simulations produced NSE values above 0.7945 & 0.824. The randomized selections that produced the depicted errors also produced 208 invalid model definitions, which caused the simulation to crash. The HSY method was applied to determine which variables were differentially modified in uniquely well-performing simulations and those responsible for crashing the simulation.

In order to determine which parameter-value pairs were responsible for (a) breaking the simulation (b) producing significantly accurate simulations, the configurations responsible for both sets of simulations were collected and assigned to three categories. Simulations producing error rates above the 5% threshold were classified as "behavioural". All other successful simulations were classified as "non-behavioural" and invalid simulation were classified as "invalid". The cumulative distribution of values was then derived for each group of values sorted into each category for each parameter. These cumulative distributions were compared calculating the Kolmogorov-Smirnov statistic on 2 samples for each parameter. The 2 sample KS test is a "two-sided test for the null hypothesis that 2 independent samples are drawn from the same continuous distribution." (Oliphant & Peterson, 2001) The cumulative distributions of parameters that produced p value < 0.001 when the behavioural & non-behavioural sets were compared are shown in Figure 3.

To get a more quantitative view of the cumulative effect of each parameter, Ridge regression (also known as Tikhonov regularization of ordinary least squares regression) was also performed, as implemented in the Linear Model module of the scikit-learn Python package (Hastie et al. 2011). The feature matrix consisted of rows for each simulation and columns containing the parameter value used during each simulation. The data was also scaled and centered, which transformed the distribution of numerical values used for each parameter to one with unit variance and a mean of 0. The results vector contained the error produced by each simulation. The coefficients of the linear model produced agreed well with the results of the HSY analysis.

Results

The following figure provides some context to how water flows through Upper Mystic Lake.

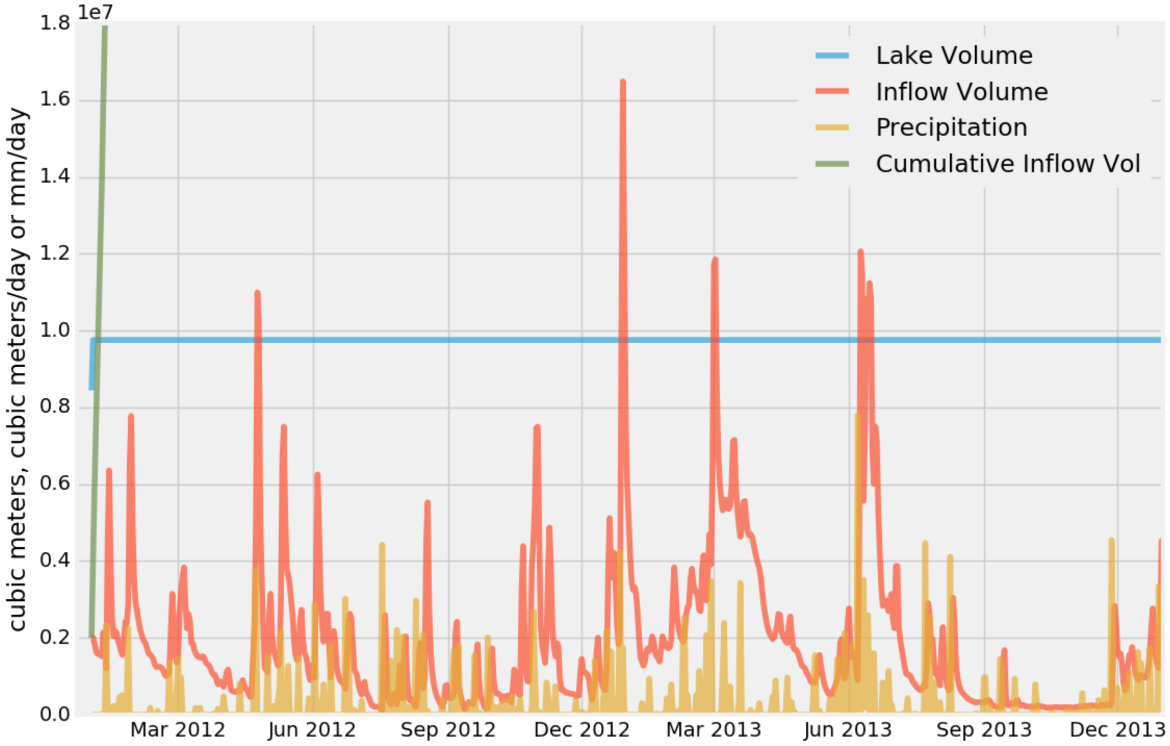


Figure 1: Water Movement Through Mystic Lake. The blue trace shows the volume contained in the water body. The red line shows how much flows into the lake per day from the Aberjona river. Notably, an adjustable weir is used to maintain the water level in the Lake, so inflow and outflow are presumed to match in the model simulation. The green line is the cumulative sum of the red line, indicating that it takes mere days for double the amount of water contained within the lake to flow through the lake. The yellow line is total daily precipitation, which exhibits strong covariance (1017.0) with inflow volume. **Note** precipitation is in units of decimeters / day and not mm / day.

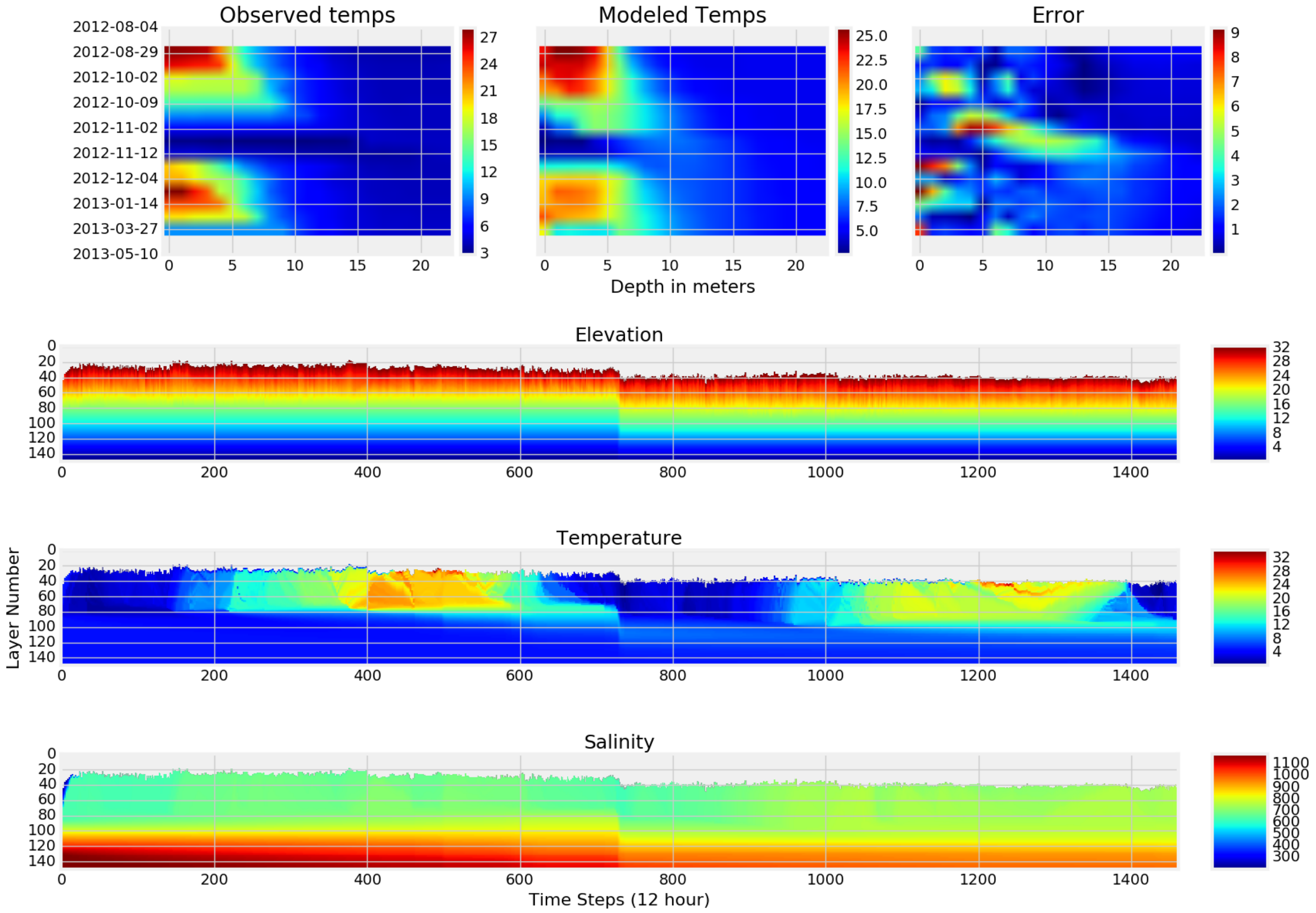
Over the time course displayed approximately 143x the standing volume of Upper Mystic Lake flows through the basin. The implied daily hydraulic loading is approximately 0.413. It is interesting to note that while nearly all observable spikes in inflow volume are preceded closely by rain events, not all rain events produce an observable bump in Inflow, nor are the magnitudes matched. However, this can be attributed to the characteristic of landscapes to act like capacitors, in a sense.

An isolated storm on July 18 & 19, 2012 that deposited a total of 0.046 meters of rain produced a daily average of 6.4 cubic meters per second of discharge before returning to baseline, whereas 10 days of steady rain between June 5-15 2013 that deposited a sum 0.1957 meters of rain, produced a total of 80.42 cubic meters / second of discharge during the same period. The difference in ratios of discharge per precipitation, which is a rough estimate of catchment area, is 35.5 km^2 for the latter and 12.1 km^2 for the former. Once the capacity for the landscape to absorb and resist flow is overmatched, the conversion from precipitation to runoff gets increasingly efficient.

Calibration

The time & depth indexed water temperatures modeled by the GLM were compared to manual measurements obtained by Dr. Preheim. When the initial values were used, had a NSE of -0.71. This was improved to a NSE of 0.79 by the greedy search algorithm from the initial parameter estimate. The randomized trials performed in preparation for HSY analysis achieved a maximum NSE value of 0.86. In more concrete terms, the initial predicted temperature values were off by about ±7.3 °C per compartment, which improved to approximately ± 1.7 °C per compartment by either calibration method. A plot of the observed water temperature distribution, as well as both sets of predictions is shown in Figure 2.

Figure 2: The best-performing simulation of the lake. The upper triptych shows a comparison of modeled and observed temperatures. The lower horizontal panels show the temperature, elevation, and salinity dynamics of individual layers during the duration of the best simulation. The influence of a storm that occurred on December 26 & 27 2012, probably compounded by snowfall on the 25th can be seen in the middle of the simulation. Although it appears as if the lake level actually fell, the model actually responded to the inflow volume by merging layers, while the top layer maintains an elevation of ~32.3 meters. The effect of the storm on the temperature distribution was profound however as the diminishing middle region that were maintaining a temperature > 10 C in the face of colder inflows from above, abruptly disappears.

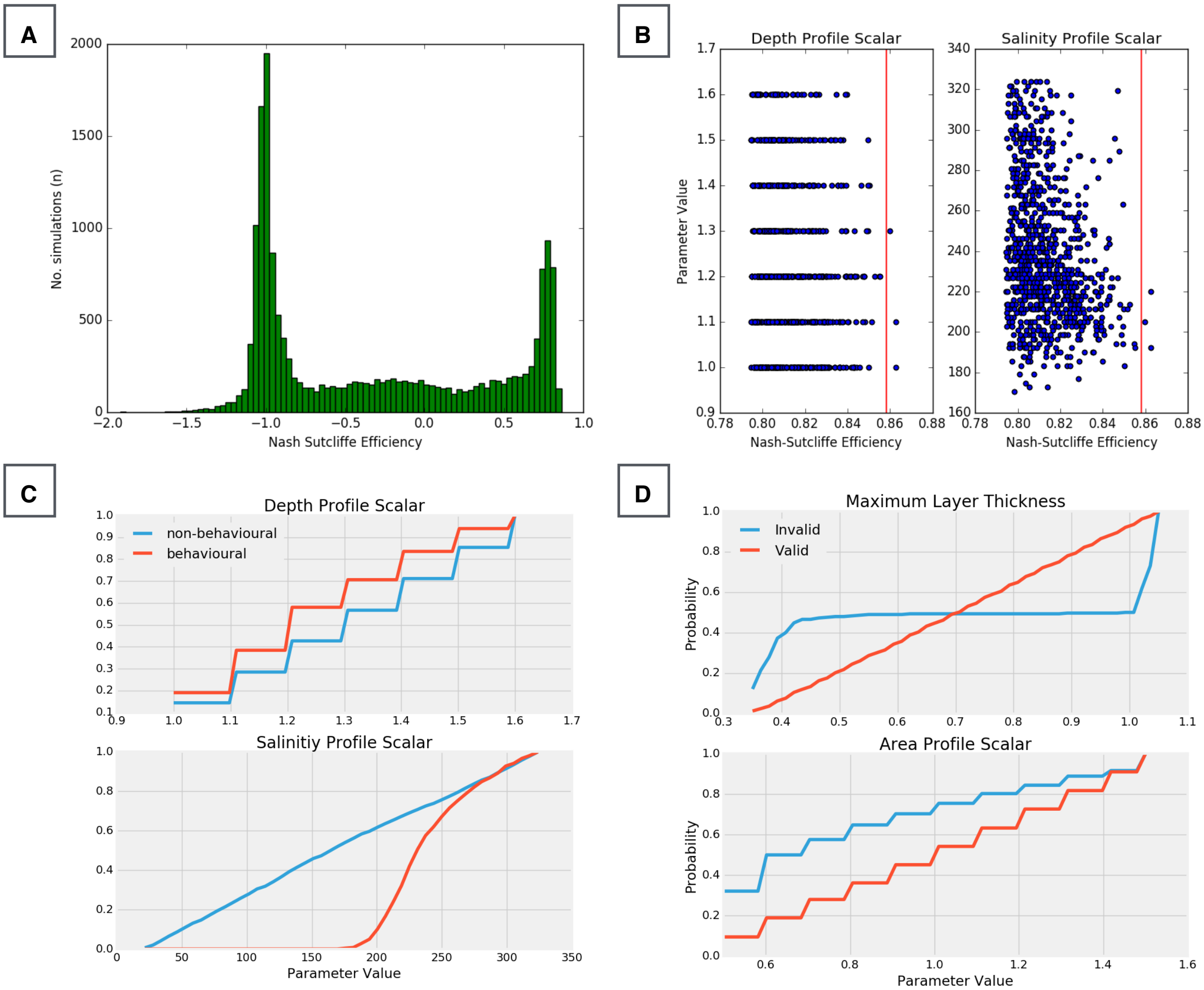


The inflow discharge volume and entrained salinity undergo high frequency variations but their first derivatives are nearly 0 (volume: 0.0476, salinity -0.0017). The steadiness of the inflow supports the maintenance of a fixed Lake volume, which is predicted by the model. Conversely, the upward migration of layers containing a salinity > 800 mg/L would not be predicted considering the steady inflow & outflow volumes, along with the steady level of inflow salinity.

Sensitivity Analysis

The parameters that showed a significantly different distribution between behavioral and non-behavioural simulations as revealed by HSY sensitivity analysis include the depth profile scalar (p= 0.009), salinity profile scalar, (p=1.2e-06), and seepage rate (p=1.8e-22). The differentially distributed parameters between invalid & valid simulations were shown to be the area profile scalar (p=0.0004) and the upper limit on layer thickness (p=7.2e-05).

Figure 3: (A)The distribution of the NSE produced by successful simulations. (B) Plots showing relations between values of the depth and salinity scalars and the error produced. The red line indicates demarcates the top 0.02% of simulations. (C & D) The cumulative distributions (CD) of differentially distributed parameters of significance between nonbehavioural & behavioural simulations (C) and valid & invalid simulations (D)



From Figure 3A, showing the histogram of NSE values, it is not discernible which among the multiple parameters is causing error values to cluster at -1.0 and 0.8, but it is evident that the effect is incremental. The previous figure shows that the non-behavioural cumulative distributions tend to be uniform. Figure 3C demonstrates that the non-behavioural CDs rise by even increments between subsequent parameter values. The Depth Profile Scalar rises ticks up steadily by ~14% and the Salinity Profile Scalar by ~1.5%. By contrast, the behavioural CDs for the Depth Profile Scalar rises by decreasing amounts, starting at 19% and falling ultimately to 6%, showing a greater proportion of values distributed toward the lower end.

The non-behavioural CD for Salinity Profile Scalar displays the opposite skew. The cumulative probability is 0% up to 169x, where it rises at an increasing rate until 235x where it begins to slow. Figure 3D shows a pair of less subtle relationships. The values for Maximum Layer Thickness in invalid simulations shows that the bulk of these values are < 0.45 or >1.0, indicating that there is a distinct acceptable zone for a given morphometry. The lower plots of Figure 3D shows another way in which the simulation fails, i.e. when smaller area profiles are used. It is most likely that particular combinations of these parameters yield unsuccessful simulations, because all of the possible values used in unsuccessful simulations were also used in successful simulations. It is also likely the malfunction is related to the settings for minimum layer volume & maximum number of layers. The latter was not subjected to the sensitivity analysis and was fixed at 200, while the former was only assigned a parameter value space consisting of 4 levels. The fact that the sensitivity analysis did not indicate minimum layer volume as a differentially distributed parameter indicates that the resolution of the test may be associated with the number of levels that a parameter can take. Figure 3B shows that the two best performing lakes show nearly identical NSE values despite the fact that one is 10% deeper than the other has nearly 15% more “salinity” at every layer. These plots also show that many, many other simulations were run with equivalent salinities and depth profiles. however mis-calibration of other parameters diminished their performance.

Ridge Regression

Ridge regression was used to assign coefficients to each parameter in accordance with their relative contribution to a given simulation’s raw error value. The coefficients with the greatest absolute value among the 36 tested are as follows with the actual coefficient in parentheses: salinity profile scalar (51.2), seepage rate (-37.3), inflow stream half angle (3.3), depth profile scalar (-2.7), long wave data scalar (2.6), wind speed scalar (-2.2), area profile scalar (1.3), and short wave data scalar (0.9).

Discussion

Simulation Behavior

The ability for the General Lake Model to simulate the physical movement of water, in the form of layers, throughout the body of a lake is most sensitive to a subset of variables revealed during the sensitivity analysis. The inflow, outflow, and seepage rates must be carefully selected when performing simulations as even a 10% decrease in the inflow, without a compensatory deduction in the outflow volume will cause the Lake to lose nearly half its volume over the two year simulation. Similarly, a seepage rate of as little as 0.1 m/day across the lake bed will drain the lake bed within a few months. To simulate the dynamic movement of water in Upper Mystic Lake, the GLM utilizes “layers” as unit control volumes. As such, they are assigned homogenous properties such as uniform temperature and concentrations, that are assigned at each time point through the basic process of budget analysis. After the initial condition is assigned, a given property is updated according to losses to adjacent layers, generative sources within a layer, and degeneration within a layer. The GLM begins the simulation of Upper Mystic Lake with 96 layers, continues with an average of 105 layers, but never exceeds 110. The volume of each layer depends on its elevation, with higher volumes appearing toward the surface. Groups of layers are then aggregated into bins at 1 meter depth increments after the simulation completes. The thickness of each layer is decided by the internal mechanics of the model, but they tend toward the lower limit set during configuration, with very few (<10%) appearing in the higher half of the range (0.25 - 0.455).

Heat & Mixing in Lakes

There are 4 types of forces that move fluid and entrained particles throughout the body of a lake. The two simple types include *advection*, which is the passive flow of bulk fluids, and *settling*, which is the vertical motion of particles within a fluid acting under the force of gravity. *Convection* is formally defined as the transfer of latent heat and results in vertical motions induced by gravity in a fluid heated from below or cooled from above. *Diffusion* is the final and smallest scale process. It can be divided into two types, molecular diffusion i.e. the random movement of particles by molecular forces and turbulent diffusion, which is the result of small-scale turbulent flows.

We can set aside diffusive forces, as the body of the lake is far too large for them to make a significant effect. We are left with the primary drivers of differential nutrient distributions throughout a lake: convection, advection, and settling. The General Lake Model is an effective engine for modeling all three of these. At present only advection & convection are simulated as the movement of particulate matter requires the Water Quality module of the GLM to be enabled, which was beyond the scope of this portion of the project.

To understand the dynamics of mixing, one must understand how thermal energy enters, exits, and distributes itself throughout a lake. An excellent source that was used to help interpret the results of the sensitivity analysis is the relevant book chapter listed in the references by Imboden & Wuest.

The largest source of thermal energy transmitted into a lake includes solar shortwave radiation, infrared radiation from the sky, which includes emission from sources other than the sun. The short and long wave radiation scalars were flagged as significant contributors to the particular distribution of heat throughout the body of the lake. These effect of these variables are modulated by cloud cover and geoposition, as the angle at which radiation enters the lake is different at different latitudes and is inhibited by cloud cover. These parameters were not indicated by the sensitivity analysis for two reasons. The first reason is that the radiation data was provided and did not need to be simulated in a process requiring geo-position as an input. The second reason is that the model may be improperly configured in some way to ignore cloud data altogether, as removing this from the meteorological database file did not have any effect on the performance metric.

Lakes also lose thermal energy from infrared emissions and evaporation. Infrared emission is modulated by differences in air temperature, relative humidity, and cloud cover. Higher air temperatures lead to greater moisture content in the air and increased scattering and deflection of infrared radiation. These conditions also cause a reduction in evaporation. The overall effect of these two heat sinks are minor and that is likely why they did not turn up in the sensitivity analysis.

The influence of inflow/outflow volumes can add or subtract heat from the body of a lake depending on the circumstance and/or season. This influence plays out on multiple levels. The mechanical energy of flowing water is translated into thermal energy by internal friction, as the current of the inflow begins to dissipate into the more placid basin volume. In addition, temperature fluxes occur more rapidly in shallow river beds and in hotter seasons, warmer inflow can directly deposit thermal energy into the body of a lake. In colder seasons, as exhibited in Figure 2B, colder water can be deposited onto warmer submerged strata. This can create a positive buoyancy flux, in which a parcel of water with higher density is deposited atop a warmer, less dense parcel, which is a precondition for convective mixing, which will be discussed below.

The effect of wind-driven force on mixing mechanisms are two-fold. It creates stress on the surface of a lake in lakes, which is roughly constant across the lake surface. This modulates the rate of latent heat transfer between the air and the lake surface. It also directly converted into mechanical energy, albeit at quite low efficiencies. According to Denman and Miyake, only about 1 to 2% of the wind power applied across the surface is transferred into the water for mixing in the upper layer. Despite this seemingly small effect, the wind scalar multiplier was considered significant by the regression analysis.

The primary factors responsible for producing low-error simulations of the temperature distribution of Upper Mystic Lake were salinity and morphometry, according to both the sensitivity & regression analysis. The morphometry parameters include a set of parameters describing the inflow stream including half angle, slope and drag. Others the seepage rate, depth profile scalar, and area profile scalar. The latter two affect the outcome of the simulation for obvious reasons, however the seepage rate is an interesting case. All preliminary simulations that included a non-zero seepage rate exhibited quite large error values. Upon inspection, it appeared as if the force responsible for pulling volume out through the bottom was transmitted vertically and leading to vertical downward migration of heat generated at the lake surface. All behavioural simulations were configured to have a seepage rate of 0. It is almost as if the entire body of the lake is treated like an outlet. Incidentally, the outlet elevation parameter also strongly effects the outcome but were not included in the analysis, as it was already determined to reflect the actual outlet elevation. All of the parameters described in this paragraph are what dictate the shape of the fluid velocity field throughout the body of the lake. They determine which layers are affected by inflowing and draining volume. They also dictate the level at which stratification takes place.

The focus of this paper was to try to assess which parameters effect how water moves vertically in the lake, but it could have been rephrased to instead focus on which parameters effect the height of lake stratification. The single greatest contributor to a high performance metric according to the regression analysis was the scalar applied to the initial salinity profile in the lake. Although it is never stated explicitly, the “salinity” parameter is used to simulate the actions of all dissolved solids, which move vertically according to the density of the entraining fluid. According to Wuest & Imboden, salinity values observed in fresh & saltwater lakes varies greatly from 0 to saturation.

In simulations where this array was set too low, the density profile of the lake was quite close to uniform, which allowed parcels of water to freely change vertical positions and transmit heat effectively throughout the lake early in the simulation. Later in the simulation, salinity brought by the inflow induced the formation of isopyncals. The resulting simulation produced two very different years in a lake. When set properly, the elevated levels of salinity in the lower layers formed a sort of density-based shield restricting the effects of solar, wind, & flow-based energy to the upper layers. The reason for this is that thermal conductivity is a colligative property and it decreases with increasing salinity. Ultimately, it was necessary to find the correct point where initial salinity values were matched to the level of inflow volume, inflow salinity, so that the distribution of dissolved species rapidly reached the steady state stratification height matched to the observed thermal profile.

Limitations & Next Steps

The GLM is a 1D model and therefore any inhomogeneities observed within layers are lost. This issue is compounded by the asymmetry of modeled data points (438,600) to observed data points (345). The inordinate value given to individual measurements, made from a somewhat arbitrary point on the surface of the lake can lead to miscalibration that plays out along a longer time scale. This is partially evident by the increase in salinity in the upper layers of the lake at later time points. One way of avoiding this is by running longer simulations, however the data available only covers the simulated period.

A distinct shortcoming of the pair of techniques used to calibrate the model is that neither can effectively used in isolation, but it isn’t clear from the presented narrative how they should be coupled. The first method, the path dependent grid search, is useful for working out which parameters need to be adjusted in relation to other set parameter values. In a sense, it can be use to untangle knots of among multiple conflicting parameter settings. Once these subsets of interacting are set in relation to one another, the remaining constellations of untethered parameters can then be effectively tweaked in bulk using running batches of random parameter value selection, followed by regression. Between each batch of random selection, the parameters that are indicated to either significantly add or subtract to the performance metrics can be set accordingly and removed from parameter value space, which should speed up each subsequent rounds of random selection.

In preparing this document, a conversion error between cubic feet per second and megaliters per day was discovered. In my effort to re-calibrate the model, it was clear that there is an upper ceiling on performance that cannot be overcome by randomization alone. As these efforts continue, the relationship between how and when calibration should follow a path and when randomization is beneficial will become clear. Once re-calibrated, further efforts in this project will be directed toward utilizing the Water Quality module of the GLM and use the AED2 engine to investigate the effects of rainfall nutrient deposition and the vertical distribution of species including oxygen, nitrogen, phosphorus, silica, organic matter of various forms, phytoplankton, zooplankton, and pathogens.

References

Jones E, Oliphant E, Peterson P, et al. SciPy: Open Source Scientific Tools for Python, 2001-, http://www.scipy.org/ [Online; accessed 2016-08-24].

Preheim, Sarah P., et al. "Surveys, simulation and single-cell assays relate function and phylogeny in a lake ecosystem." Nature Microbiology 1 (2016): 16130.

Hunter, K. S., Wang, Y., & Van Cappellen, P. (1998). Kinetic modeling of microbially-driven redox chemistry of subsurface environments: coupling transport, microbial metabolism and geochemistry. Journal of hydrology, 209(1), 53-80.

Trevor J.. Hastie, Robert John Tibshirani, and Jerome H. Friedman. The elements of statistical learning: data mining, inference, and prediction. Springer, 2011.

Denman, K. L., & Miyake, M. (1973). Behavior of the mean wind, the drag coefficient, and the wave field in the open ocean. Journal of Geophysical Research, 78(12), 1917-1931.

Imboden, Dieter M., and Alfred Wüest. "Mixing mechanisms in lakes." Physics and chemistry of lakes. Springer Berlin Heidelberg, 1995. 83-138.

Read, Jordan S., et al. "Simulating 2368 temperate lakes reveals weak coherence in stratification phenology." Ecological Modelling 291 (2014): 142-150.

Bruggeman, Jorn, and Karsten Bolding. "A general framework for aquatic biogeochemical models." Environmental Modelling & Software 61 (2014): 249-265.

Hipsey, M. R., Bruce, L. C., & Hamilton, D. P. (2013). GLM General Lake Model. Model overview and user information. The University of Western Australia Technical Manual, Perth, Australia.

Beven, Keith J. Rainfall-runoff modelling: the primer. John Wiley & Sons, 2011.

Appendix I

This table shows all the parameters of the General Lake Model that have a qualitative effect on the outcome in the first column. The second column shows the values selected during the best simulation and the third column provides a brief description. Data types and units are shown when provided by the model's authors

Parameter	Value Selected	Unit/Description
GLM Setup		
max_layers	300	layers
min_layer_vol	0.02	meters
min_layer_thick	0.32	meters
max_layer_thick	1.0	meters
Kw	0.6	light attenuation (1/m)
coef_mix_conv	0.1625	convective overturn
coef_wind_stir	0.6375	wind stirring
coef_mix_turb	0.315	unsteady turbulence effects
coef_mix_shear	0.675	shear production
coef_mix_KH	0.315	hypolimnetic Kelvin-Helmholtz turbulent billows
coef_mix_hyp	0.675	hypolimnetic turbulence
deep_mixing	FALSE	flag to disable deep-mixing
Morphometry		
latitude	42	degrees North
longitude	-71	degrees East
bsn_len	1012	basin length at crest (m)
bsn_wid	536	basin width at crest (m)
bsn_vals	9	number of depth points on height-area relationship
H	0., 3., 6.1, 9.1, 12.2, 15.2, 18.3, 21.3, 24.4	comma separated list of elevations (m) (must monotonically increase)
A	77,374, 148,476, 202,473, 257,819, 338,553, 397,077, 460,778, 524,803, 560,051	comma separated list of areas (m^2) (must monotonically increase)
Time		

start [string]	2012-01-01 00:00:00	nominal start date
stop [string]	2014-01-01 00:00:00	nominal stop date
dt [float]	3600.0	time step for integration (seconds)
time_zone	5.0	time zone number code
Output Parameters		
nsave [integer]	1 to 86400 (24)	values are output every nsave time steps
Initialization Parameters		
lake_depth [float]	24	initial lake depth (m)
num_depths [integer]	6	number of depths provided for initial profiles
the_depths [float]	[1, 5, 9, 13, 17, 21]	the depths of the initial profile points (m)
the_temps [float]	[4.0, 4.0, 4.0, 4.0, 4.0, 4.0]	the temperature of the initial profile points (C)
the_sals [float]	[200., 400., 600., 800., 1000., 1200.]	the salinity of the initial profile points (mg/L)
Meteorology Parameters		
met_sw [bool]	True	switch to include surface meteorological forcing
lw_type [string]	LW_IN	directional flux of longwave data (down, down & up, or net down)
rain_sw [bool]	False	include rainfall nutrient composition
atm_stab [bool]	False	account for non-neutral atmospheric stability
catch rain [bool]	False	flag that enables runoff from exposed banks of lake area
rad_mode [integer]	0	short and long wave radiation model configuration. 0 indicates daily values of solar & cloud data provided
albedo_mode [integer]	1	shortwave albedo calculation method (see pg. 9 of documentation)
cloud_mode [integer]	4	atmospheric emmisivity calculation method (see pg. 10 of documentation)
wind_factor [float]	1	wind multiplier (+/-)
rain_factor [float]	1	rain multiplier (+/-)
sw_factor [float]	1	short wave radiation multiplier (+/-)
lw_factor [float]	1	long wave radiation multiplier (+/-)
at_factor [float]	1	air temperature multiplier (+/-)

rh_factor [float]	1	relative humidity multiplier (+/-)
ce [float]	0.001	bulk aerodynamic coefficient for latent heat transfer
ch [float]	0.0011	bulk aerodynamic coefficient for sensible heat transfer
cd [float]	0.0015	bulk aerodynamic coefficient for transfer of momentum
rain_threshold [float]	0.03	rainfall amount (m) required before runoff from exposed banks
runoff_coef [float]	0.1	conversion of rainfall to runoff in exposed lake banks
Inflows		
strm_hf_angle [float]	1.1	stream half angle (degrees)
strmbd_slope [float]	3.35	streambed slope (degrees)
strmbd_drag [float]	0.008	streambed drag coefficient (+/-)
inflow_factor [float]	0.8	inflow flow rate multiplier (+/-)
coef_inf_entrain [real]	0.7	entrainment coefficient for inflows
Outflow Parameters		
flt_off_sw [bool]	FALSE	floating offtake switches
outl_elvs float]	24	outlet elevations (comma separated list)
bsn_len_outl [float]	100	basin length at outlets (m)
bsn_wid_outl [float]	274	basin width at outlets (m)
outflow_factor [float]	0.5	outflow flow rate multiplier (+/-)
seepage [bool]	FALSE	do seepage processing
seepage_rate [float]	0.0	seepage rate of water (m/day) from bottom layer

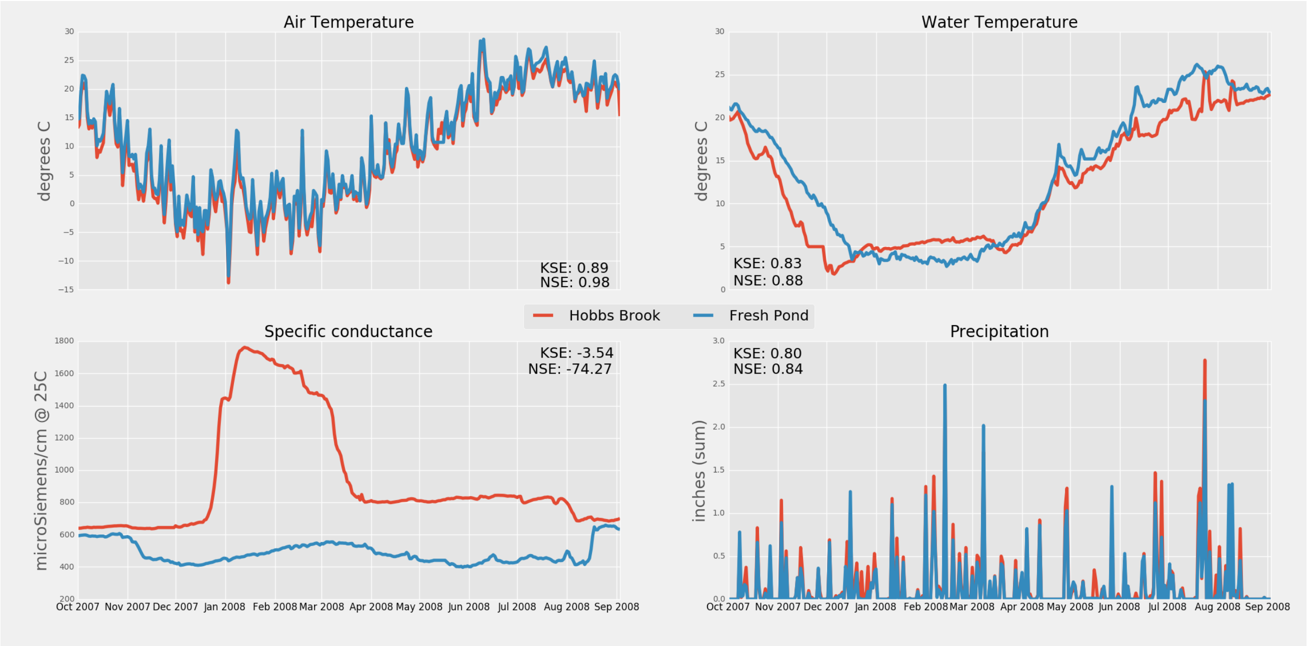
Appendix II

Parameter value space: The parameter block is specified first and signified by a "&", which is consistent with the original GLM configuration file. Each variable within that block is then listed below it on a line started with "-". The information about values is specified in the first set of brackets and the datatype in the second. If it is a boolean, true and false strings are specified without quotes in the same format as in the original configuration file. Integers and floats are specified using the syntax `[start, stop, step]`. If step is `None`, only the start and the stop values are tried. Lines that start with hash marks are ignored.

```
&glm_setup
- deep_mixing = [.true., .false.],[bool]
- Kw = [0.25, 0.75, 0.025],[float]
- min_layer_vol = [0.018, 0.022, 0.001],[float]
- min_layer_thick = [0.10, 0.34, 0.01],[float]
- max_layer_thick = [0.35, 1.05, 0.01],[float]
- coef_mix_conv = [0.0625, 0.1875, 0.00625],[float]
- coef_wind_stir = [0.115, 0.345, 0.0115],[float]
- coef_mix_shear = [0.01, 0.3, 0.01],[float]
- coef_mix_turb = [0.255, 0.765, 0.0255],[float]
- coef_mix_KH = [0.15, 0.45, 0.015],[float]
- coef_mix_hyp = [0.25, 0.75, 0.025],[float]
&morphometry
- longitude = [-161, 170, 15],[float]
- latitude = [-78, 73, 15],[float]
- bsn_len = [55.0, 2145.0, 55.0],[float]
- bsn_wid = [30.0, 1170.0, 30.0],[float]
- A = [0.5, 1.5, 0.1],[list]
- H = [1.0, 1.5, 0.1],[list]
&init_profiles
- the_temps = [0.1, 1.5, 0.1],[list]
- the_sals = [0.1, 1.5, 0.01],[list]
&outflow
- outl_elvs = [0.1, 24.1, 1.0],[float]
- seepage_rate = [0.0, 0.2, 0.05],[float]
- bsn_len_outl = [99, 1099, 50],[int]
- bsn_wid_outl = [24, 599, 25],[int]
- outflow_factor = [0.8, 1.0, 0.05],[float]
&inflow
- coef_inf_entrain = [0.0, 1, 0.1],[float]
- strmbd_drag = [0.004, 0.0320, 0.004],[float]
- strmbd_slope = [2.05, 4.0, 0.05],[float]
- strm_hf_angle = [0.05, 2.0, 0.05],[float]
- inflow_factor = [0.5, 1.0, 0.1],[float]
&meteorology
- met_sw = [.true., .false.],[bool]
- catchrain = [.true., .false.],[bool]
- rad_mode = [1, 5, None],[int]
- albedo_mode = [1, 3, 1],[int]
- cloud_mode = [1, 4, 1],[int]
- rain_threshold = [0.00, 0.05, 0.01],[float]
- runoff_coef = [0.0, 0.7, 0.05],[float]
- wind_factor = [-1.5, 1.5, 0.05],[float]
- rain_factor = [-1.5, 1.5, 0.05],[float]
- at_factor = [-1.5, 1.5, 0.05],[float]
- rh_factor = [-1.5, 1.5, 0.05],[float]
- sw_factor = [-1.5, 1.5, 0.05],[float]
- lw_factor = [-1.5, 1.5, 0.05],[float]
- cd = [0.00065, 0.00195, 6.5e-05],[float]
- ce = [0.00065, 0.00195, 6.5e-05],[float]
- ch = [0.00065, 0.00195, 6.5e-05],[float]
```

Supplementary Information

Figure S1 : Available USGS Data for two nearby sites (Fresh Pond & Hobbs Brook): This data was plotted in order to get a sense of regional variability in precipitation, water & air temperature, and specific conductance. As is plainly evident, conductance is a site specific parameter. A zoomed in view of the plot on the upper left is presented further below with some important differences. Data from 2012-2013 for Fresh Pond is presented along with surface temperature measurements from the body of Mystic Lake from the same year and the identical info from Hobbs Brook in 2007-08 (shown in the upper panel). This provides a sense of the variability between years and the comparability of Mystic Lake with nearby water bodies.



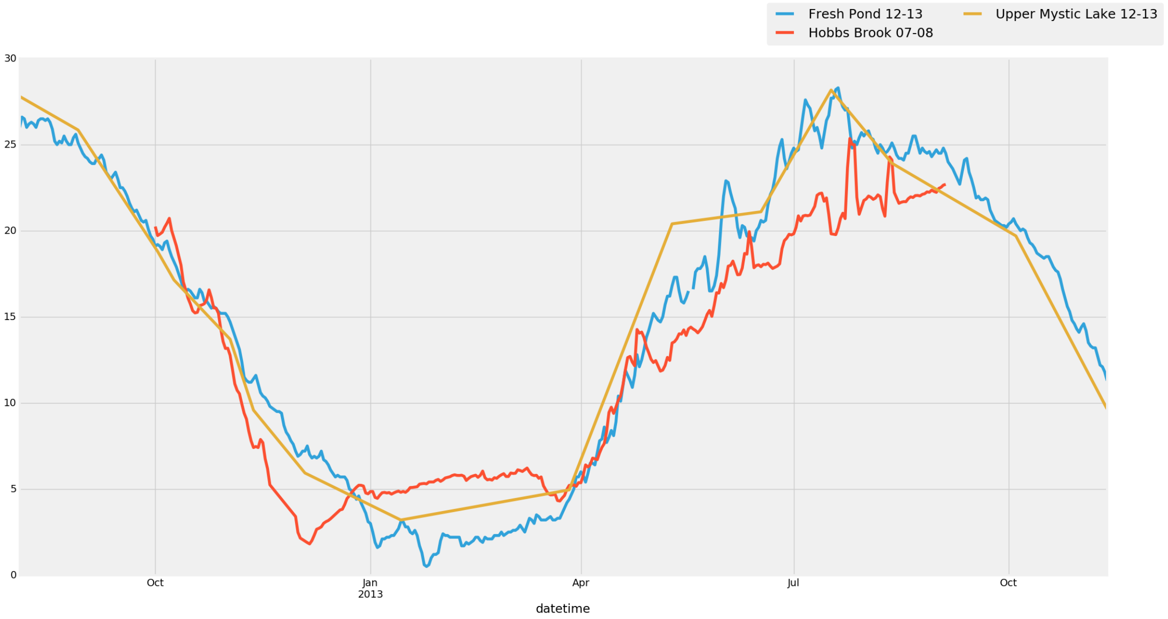


Figure S2 : Comparison of Air/Skin Surface Temperature: NOAA NEIS, CERES, and USGS Water Data Archives reveal slight location-based variations in temperature

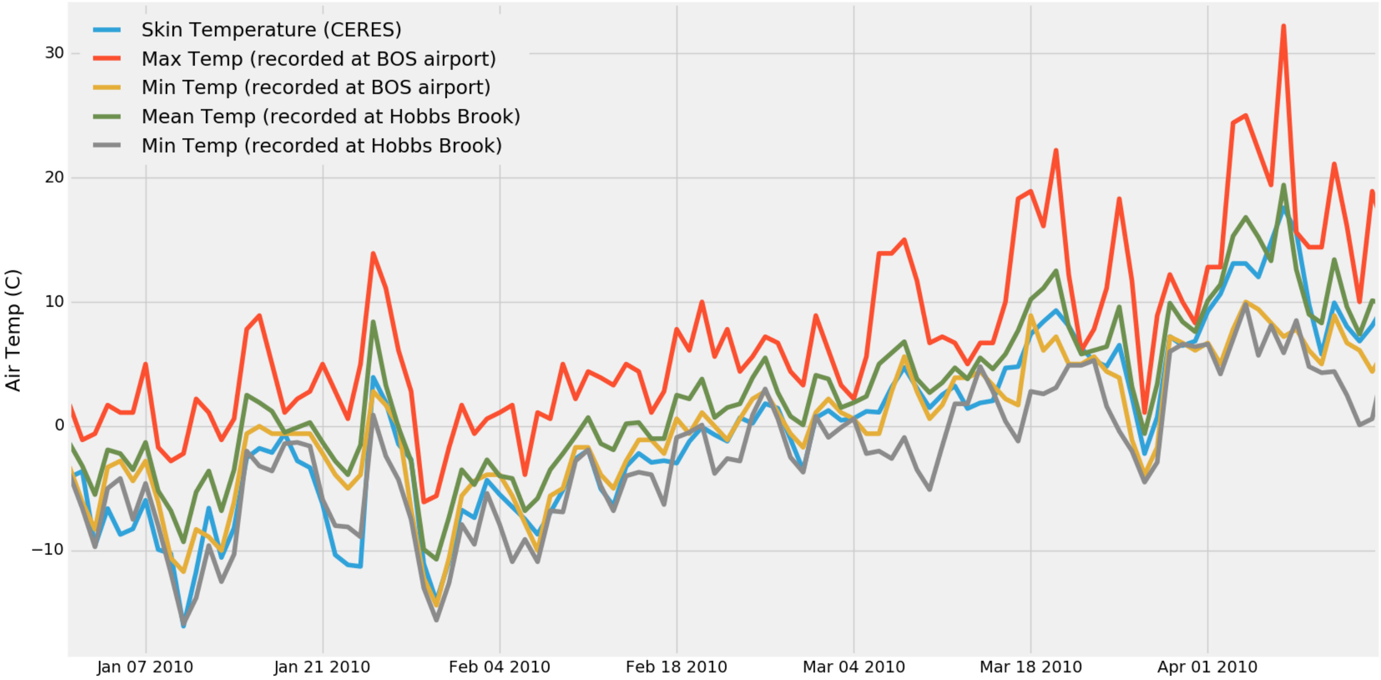


Figure S3: Location/Extent of Relevant Locations: Inflow River (H, Abjerjona), Upper Mystic Lake (A), Outflow River (F, Mystic River), two nearby water bodies used for comparison (B, Fresh Pond & C, Hobbs Brook aka Cambridge Reservoir) and geopositional locus for CERES satellite data (G) and NOAA/NCEI weather data (D, Boston Logan International Airport)

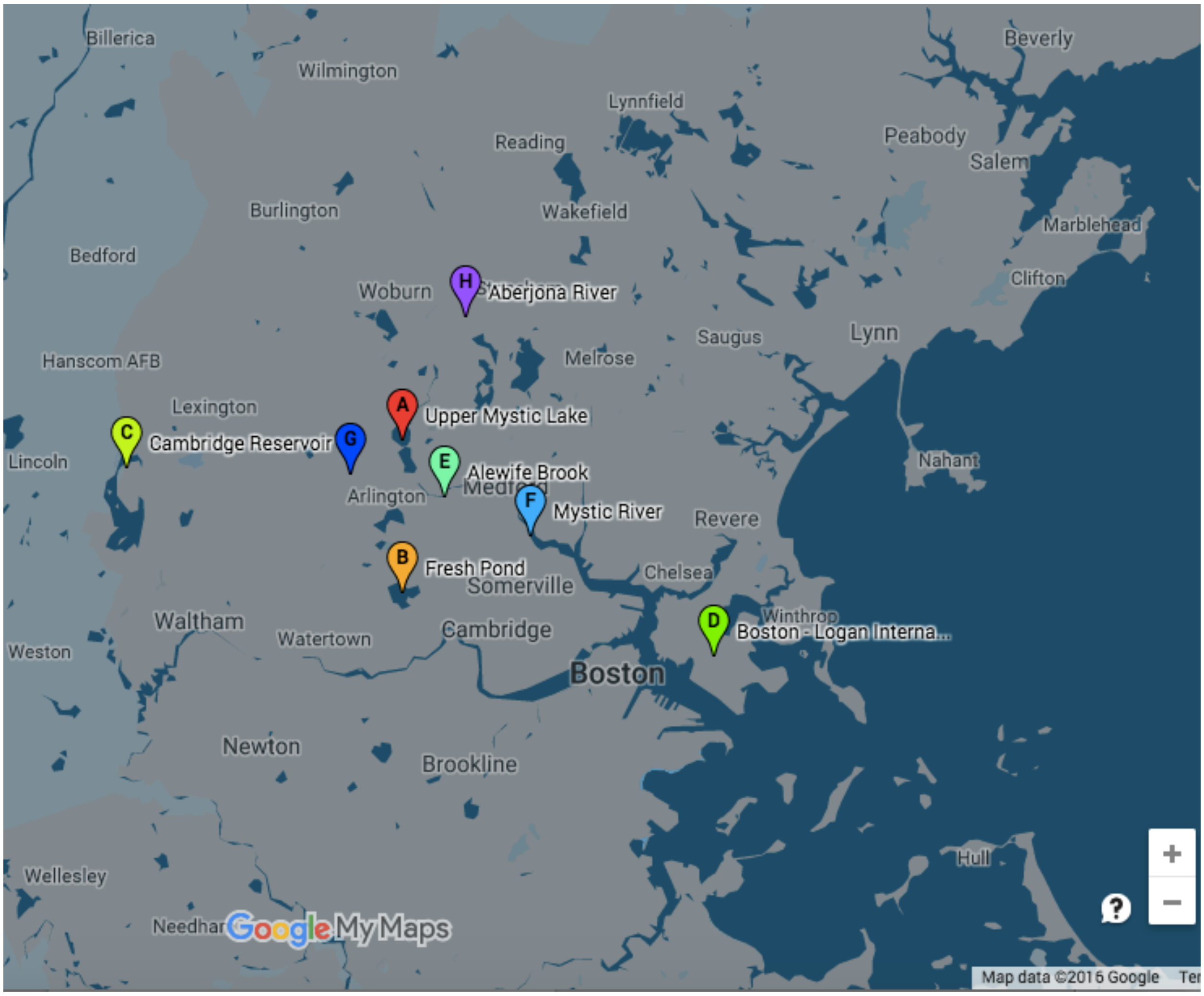
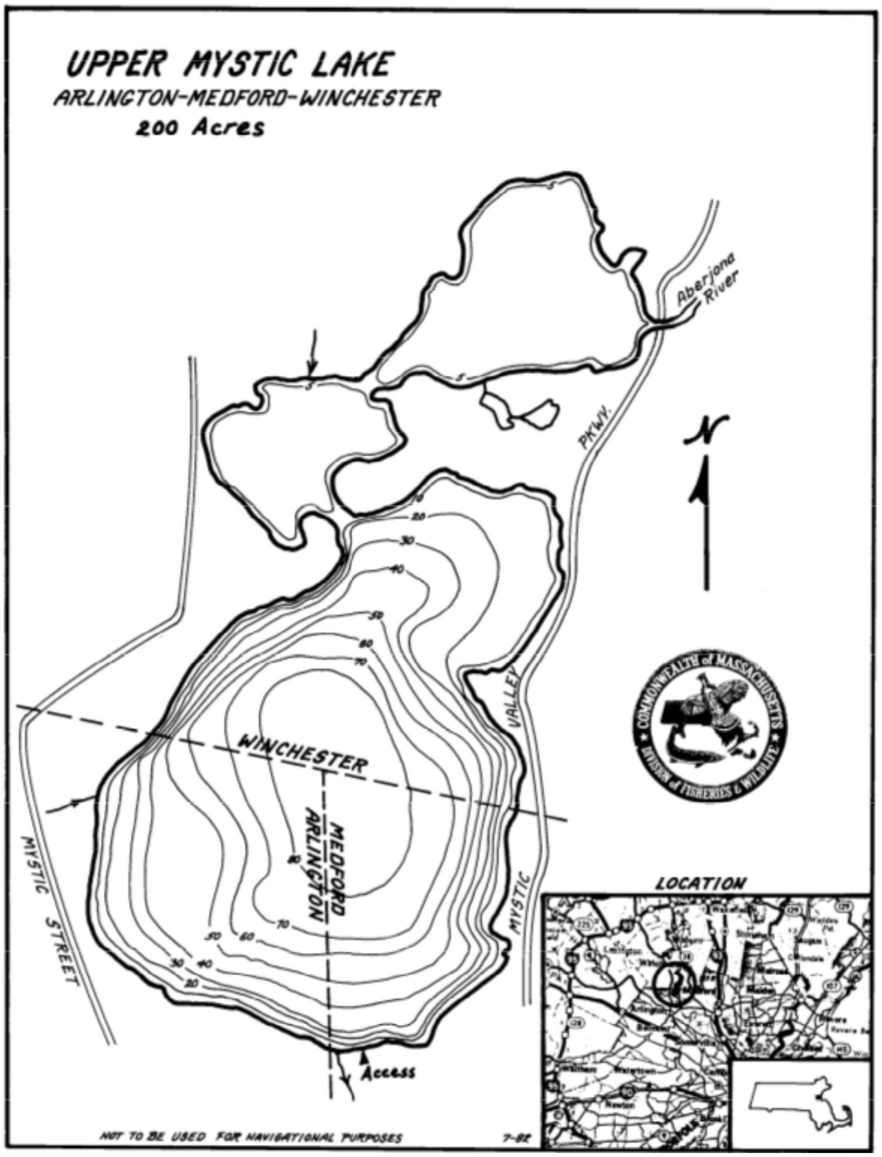


Figure S4: Hypsographic contours: Depth-area relationship was derived from this image.


UPPER MYSTIC LAKE
ARLINGTON-MEDFORD-WINCHESTER
200 Acres
Aberjona River
PKWY.
VALLEY
MYSTIC
MYSTIC STREET
WINCHESTER
MEDFORD
ARLINGTON
5
10
20
30
40
50
60
70
80
Access
COMMONWEALTH of MASSACHUSETTS
DIVISION of FISHERIES & WILDLIFE
LOCATION
NOT TO BE USED FOR NAVIGATIONAL PURPOSES
7-82