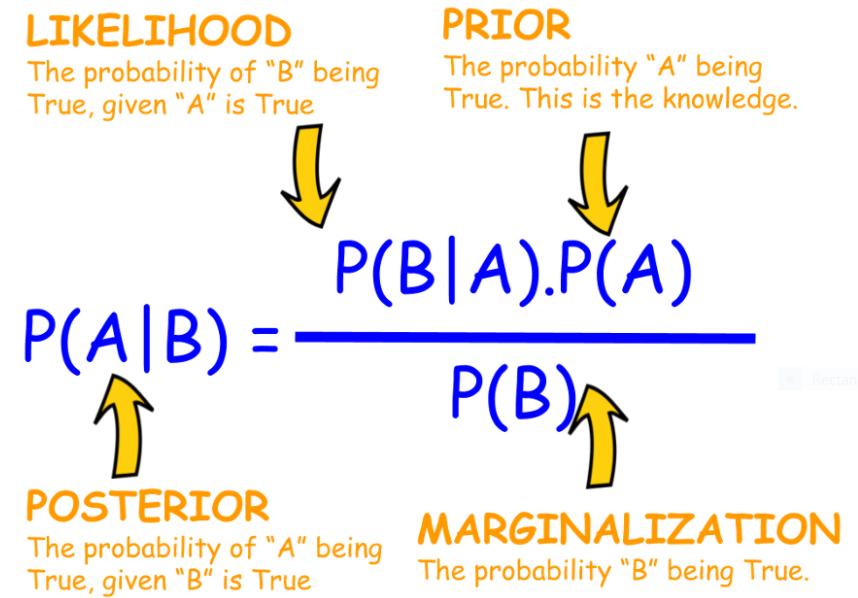
**L2SWBM Overview**

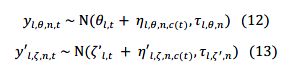
To start we need to distinguish between frequentist and Bayesian statistics. Frequentist statistics is what most people are familiar with. If we were to use a frequentists approach to the L2MWBS model, we’d use all the data to determine the probability of a given value occurring. In Bayesian statistics, we incorporate known information to determine a likely outcome. The information we know about the process is the prior information and in this instance expert opinion is built in as well. So in the current configuration of the model, Smith and Gronewold (2018) looked at historic data from 1950 to 2008 and used these data to estimate the prior distributions for the individual water balance components. The methodology for this is the same as would be carried out in a frequentists approach (i.e. they looked at histograms of the data to determine the most likely distributions).

Bayesian statistics are carried out through the application of Bayes Theorem:

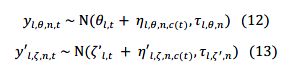


i

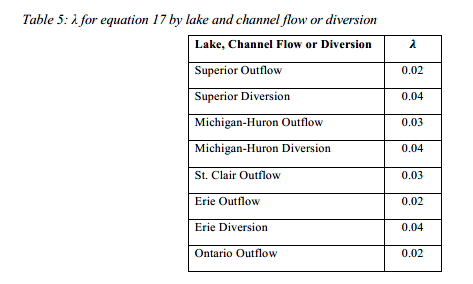
In this example, , is the water balance component and is the historical data. In the case of the L2SWBM, both and are historical data but they have different periods of record. is the probability of the prior distribution. In the L2MWBS model this is determined through fitting appropriate probability distributions to historical data from 1950 to 2008. is the likelihood, meaning the probability that an observation will occur given what we’ve seen in the past. For the L2MWBS model the likelihood is specified as (Smith and Gronewold, 2018):



Everything denoted by a prime is for Lake St. Clair. Water balance calculations for Lake St. Clair were carried out differently due it having a smaller surface area and there being more uncertainty in the components of its water balance. My explanation will focus on equation 12 for the sake of brevity and because the equations were formulated in the same manner. is the monthly observation for lake , for water balance component , for data source , and month . It is assumed/known that there is bias in these estimates, which is denoted by , which is the bias for lake , for for water balance component , for data source , and is the calendar month. I believe that and are equivalent. As you can see above in equations 12 and 13 the normal distribution is selected as the likelihood function, which has two parameters (mean and standard deviation):



Therefore, bias is built into the mean and the standard deviation is represented by precision estimates. is the true estimate of the water balance parameter in the absence of bias. The precision (equal to 1/variance) is informed from previous uncertainty analyses and expert opinions. Smith and Gronewold (2018) specific the uncertainty in their Table 5 (included below). You can see these as input parameters in the config.csv file. So these are expertly informed.



Back to Bayes Theorem. We’re interested in finding the posterior distribution, which it is the probability distribution that describes uncertainty in the water balance components after having considered what we know about the data. So we know 1) the approximate distribution of the data from the historical period of record (the prior) and 2) the likelihood of those historic estimates occurring based on historical data from a different period of record (the more recent period of record – the config.csv file is set up to run the last 10 years of monthly data) and expert opinion. is also referred to as “the evidence” because it’s the evidence that the data were created by the proposed model. It also serves the purpose of normalizing. There are rarely closed formed solutions to this component of Bayes theorem, therefore, it’s generally reduced to:

Unormalized posterior = likelihood x unormalized prior

Which is a simplified form of Bayes theorem and the one that the L2SWBM model is based upon.

The run time of the model is the result of the Monte Carlo Markov Chain (MCMC) simulations. Each water balance component is being simulated with three chains and each chain is run 100,000 times in the current configuration of the model.

The steps of the simulation are as follows:

1. Start the chain with a randomly (but informed) started point. This means that a realistic value of the posterior distribution is selected to start the simulation.
2. Because we don’t know if this random starting point is a good choice, it will take some time for the chain to become stationary or converge. The time to achieve this is referred to as the burn-in period and Smith and Gronewold (2018) omit the first half of the chains to account for this.
3. The Markov aspect of the chains comes from the dependence of each simulation on the one that precedes it (which I’ve discussed in greater detail below). Because of this dependence, there is inherent serial correlation in the data, which is undesirable. To account for this, the data are randomly thinned. Therefore, the remaining 500,000 simulated points are thinned out until only 1000 remain.
4. Since three chains are simultaneously run, we end up with 3000 simulated point for each water balance component.

The steps of what happen when each randomly selected point from the posterior distribution are selected are as follows:

1. An informed but random value of the water balance component is selected.
2. We then determine another plausible value of the posterior distribution and “propose a jump”.
3. We can calculate the probability of the likelihood of occurrence of these points and the probability of the estimates from the prior distribution.
4. By multiplying the two aforementioned probabilities, we arrive at the probability of that value in the posterior distribution.
5. The MCMC algorithm then has to decide if that point is a good place to be or if it should jump to the proposed point.
6. The probability that the proposal is accepted is:

Probability of Acceptance = Probability of the Proposal/Current Probability

1. If the probability of the proposal is greater than the current probability, the chain will move there (P > 1). As example is the probability of the proposal is 0.2 and the current probability is 0.4, there’s a 50% change the model will jump to that state.

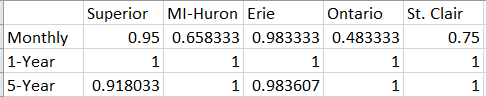
These steps continue until the desired number of iterations have been realized. As you can see, the model will eventually get to a state is stationary and will be less likely to move away from that state based on probabilities.

There are different types of sampling that produce the proposal. Some are much more sophisticated than others. Smith and Gronewold (2018) use Gibbs sampling, which is common and robust and easily carried out in R.

Now pulling this all together in terms of the water balance. Data are available from 1950 to the present. In the current configuration of the model data from 1950 to 2008 and used to inform the prior distribution and aren’t used in the water balance. We can consider this previous information on the expected distribution of those parameters. Any data after December 2008 can be used to calculate the water balance. The water balance is related to the Bayesian MCMC simulation through the likelihood function. We have monthly observations but we don’t know if they’re true due to bias, therefore, the model provides the correct estimate of the water balance component based on simulated data.

**Additional Notes:**

* L2SWBM is very good at closing the 1-year and 5-year water balance but can be poor at a monthly time scale. When you run the model, there’s a ‘ModelName\_Closure.csv’ file that’s generated is requested in the config.csv file. This is from October’s run:



I think we have to be critical as to how we use these results on a monthly time scale. This might vary from month to moth though.

**Next Steps:**

In order to calibrate the model we first have to find the minimum number of iterations needed for convergence. The model is currently running 100,000 but it’s possible we could run less without a loss of accuracy. If André has some insight on what he’s tried to change, it would be helpful. Also any changes that Frank and Lauren made that decreased the uncertainty would be helpful. From there, we can focus on individual parameters and see if we can reduce the uncertainty in the model with parameter selection.