# Eighth Montreal Industrial Problem Solving Workshop - Rio Tinto Report

D. Jovmir[1], N. Ayi[2,3], A. Poterie[4], C. Budd[1], S. H. Jun[5], K. A. Alahassa[1], S. Amraoui[6], S. Ibrahim[1], T. Y. Lee[1], C.P. Liou[1], C. Poissant[7], V. Rochon Montplaisir[1], L. Sarrazin-Mc Cann[1], P. Duchesne[1], R. Arsenault[8], and M. Latraverse [8]

[1]Université de Montreal, Département de mathématiques et statistique, Montréal, Québec
[2]Sorbonne Universités, UPMC Univ Paris 06, UMR 7598, Laboratoire Jacques-Louis Lions, F-75005, Paris, France
[3]CNRS, UMR 7598, Laboratoire Jacques-Louis Lions, F-75005, Paris, France
[4]INSA-IRMAR, 20 Avenue des Buttes de Coesmes, 35708 Rennes, France
[5]University of British Columbia, Statistics Department, Vancouver, British Columbia
[6]University of Nice, CNRS, LJAD, Nice, France
[7]École Polytechnique de Montréal, Département de mathématiques et de génie industriel, Montréal, Québec
[8]Rio Tinto

23 janvier 2018

**Résumé**

Rio Tinto uses a complex hydrological model to make ensemble predictions (ESP) for expected freshet volumes that are vital to planning the management of their hydro-electrical power plants. However, the ESP predictions show an under-dispersion problem that is apparent in Talagrand histograms used for evaluating model performance. This problem has been successfully solved for the winter season [1] but the summer season presents unique challenges. During the workshop, we proposed four different approaches to correcting the under-dispersion problem in summer : optimization of the initial state in summer , transfer function models, Gaussian processes and one dimensional modeling.

# 1 Introduction

Rio-Tinto operates six hydro-electrical dams throughout the Saguenay-Lac St Jean region that provide 90% of energy used at their aluminum smelting plants. Since hydro-power generation is a function of flow rate through the turbines, which in turn depends on water levels in the reservoirs, some planning must go into water level management. Part of day to day operations consists of predicting inflows (which integrate to expected water-levels in the reservoirs) for the next 14 days. This allows to plan for optimal outflows in advance in order to maximize electricity production as well as minimize inconveniences for neighboring communities and other users of the reservoirs.

Water flowing into a reservoir is a function of geography of the catchments and soil properties as well as weather and precipitation, and thus can't be predicted exactly.

## 1.1 Hydrological model

Rio-Tinto uses the CEQUEAU model (a complex time-dependent hydrological model) to estimate resulting inflows that nevertheless requires daily manual adjustments.

The model takes temperature and precipitation forecasts as inputs ($I_t$). It then uses initial conditions ($x_t$) as well as other parameters and information about the geography of the terrain ($\theta$) to derive 2 additional state variables ; $UW_t$ : the water level in the aquifer and $SW_t$ : the amount of water found in the soil. During the winter and spring season, the dominant initial condition is the amount of snow accumulated on the ground ($SN_t$). This variable is measured at different sites throughout the winter but measurement errors can be substantial. The observed inflows ($Q_t$) can then be modeled as :

$$Q_t = M(x_t, \theta, I_t) + \epsilon_t \tag{1.1}$$

where M(.) represents the estimated outputs calculated by the hydro model and $\epsilon$ is an error term stemming from uncertainty in weather observations, initial conditions as well as modeling errors.

In reality, the observed inflows rarely match the estimated inflows. Before making predictions for the following 2 weeks, an experienced technician has to adjust the values of initial conditions until the estimated inflows match observed inflows. This adjustment ensures that initial conditions are as close to reality as possible before making predictions for a new cycle. However, the method by which these adjustments are made is mostly based on the intuition of the individual technician and thus can't be replicated.

Once the initial conditions of the hydrological model are adjusted, predictions can be made. Since long range weather forecasts can be inaccurate, Rio Tinto uses ensemble predictions. Historical weather data, available for the prediction period since 1954, is fed into the model, thus producing 63 different predicted scenarios for the upcoming weeks. Since a fair amount of weather data is available, they are confident that their predictions will cover most possible outcomes. The next step in the decision process takes the different scenarios as inputs and makes an optimal decision on how to manage water levels to maximize power output. However, it is important that the input scenarios be equally likely. This assumption can be tested (in retrospect) using a Talagrand histogram [5] (or PIT graph).

## 1.2 Talagrand Histograms

To test the assumption of equal probability for each member of the ensemble predictions, the sum of inflows over the 14 days of the prediction period is calculated and compared with the observed water accumulation during this time. This comparison is made for each year for which weather data is available and the quantile in which the observed value lies is recorded. The observation should fall in all quantiles of the predicted scenarios with equal probability. A histogram of the observed quantiles should appear flat (thus following a uniform distribution). The resulting histogram is called the a Talagrand Histogram [5] or PIT graph, and it is the tool used by Rio Tinto to asses the performance of their hydrological model.

## 1.3 Problem

It has been observed that the ensemble predictions output by the model appear to be consistently under-dispersed. This translates into predictions that often fall above or below the observed inflows and manifest as a concave PIT graph (more density on the edges and less in the middle). The problem is further complicated by the fact the model response to initial conditions is season specific. During the spring, snow melt is largely responsible for all inflows into the reservoir. However, during the summer, all initial conditions have a considerable effect on inflows and the initial conditions themselves vary much more quickly according to weather conditions.

## 1.4 Solutions

Rio Tinto hydrologists have been successful at solving this problem in an elegant way for the winter/spring melt season [1]. Their method considers measurement errors that occur when setting the snowpack ($SN_t$) variable as part of initial conditions. With the help of historical information, the distribution of the $SN_t$ variable is approximated. The distribution is then resampled a number of times (let's say 10) and ensemble predictions are then produced as before for each of 10 sets of initial conditions. This method then yields ($63 \times 10$) 630 projected outcomes that are then input into the decision process. The predictions made using this method are found to be more adequately dispersed than the previous approach, that didn't address the variability in snowpack measurements.

However, this solution hasn't yet been succesfully adapted to summertime predictions, due to the complexity of having to consider all other initial conditions important during the summer.

In the following sections we will describe the five different approaches that we explored during the week in our attempts to correct the under-dispersion of ensemble forecasts for the summer.

In section 2 we adapt the method for adding variability to winter predictions to the summer season.
In section 3 we adopt a post-production approach, where we attempt to fix model predictions using time series methods. We estimate the bias of the predictions using a transfer function model with some or all initial conditions as explanatory variables and then apply this model to modify the ensemble predictions in the hopes that variability will be more accurately described.
In section 4 we use Gaussian processes to model the relationship between the inputs and outputs

of the CEQEAU model.
In section 5 we produce a simple one dimensional hydrological model which can then be used to test various data assimilation approaches.

# 2 Optimization in the initial state in summer

## 2.1 Explanation of the winter method

Rio Tinto uses the CEQUEAU model to produce ensemble streamflow prediction (ESP). This model estimates the freshet volume, which is the hydrological variable of interest, based on parameters $\theta$ (calibrated on the data), initial conditions $x$ and climate inputs $I$. It is usually denoted $M(x, \theta, I)$.
The problem with this model is that the ESP forecast are often under-dispersed. This problem can mainly be explained by the fact that the hydrological CEQUEAU model is purely deterministic. Thus, it does not fit the fact that, in the real world, sometimes when using similar climate and/or state variables, a certain variability can be observed. Besides, the model only uses years on record for the climate. Therefore, there is a limitation on the possible outcomes.

To correct the under-dispersion, the idea is then to find a way to reintroduce the missing variability into the ESP forecast members. This is actually the start of the method developed for the winter and introduced in the paper [1]. This method will be called the $\Delta V$ method in the following sections.
First, in winter in the sub-basin of the Lac-St-Jean watershed in central Quebec, the dominant hydrological variable related to the freshet volume is the snow water equivalent (SWE) on the catchment. Thus, this state variable will be used to reintroduce some variability.

We denote by $y$ the observed freshet volume and $t$ a particular year. Thus, $\varepsilon(x_t|\theta, I_t)$ represents the error between the observed freshet volume and freshet volume estimated with the CEQUEAU model which is denoted by $M(x_t, \theta, I_t)$. Therefore, we have the following relationship :

$$y_t = M(x_t, \theta, I_t) + \varepsilon(x_t, \theta, I_t). \tag{2.1}$$

In order to estimate the model error based on historical simulations, the former equation was modified by conditioning on the climate input and the parameter $\theta$. It leads to the equation :

$$y_t = M(x_t|\theta, I_t) + \varepsilon(x_t|\theta, I_t). \tag{2.2}$$

The method used to reintroduce the missing variability into the ESP forecast members follows the steps described below :

1. **Hindcast Step**
For each year $t$, ($t = 1954, \dots, 2014$)

   (a) we run the hydrological model with fixed parameter set $\theta$ using observed climate data $I_t$ for the required ESP duration.
   $\Rightarrow$ We obtain $M(x_t|\theta, I_t)$.

   (b) We calculate the error term $\varepsilon(x_t|\theta, I_t)$ based on equation 2.2.

   (c) We have identified previously the scalar value of the model in the vector $x$ with which we will play : the SWE. We apply a correction to SWE in order to reduce the error. We denote by DSWE the corrected value of SWE in $x$ for Delta snow water equivalent.
   $\Rightarrow$ We obtain $DSWE_t$.

   The previous process is repeated for each year $t = 1954, \ldots, 2014$ : we obtain $\{DSWE_{1954}, \ldots, DSWE_{2014}\}$. We can then model the DSWE distribution (in a parametric or non parametric way).

2. **Prediction Step**

   (a) For each year, we pick random sample from the DSWE distribution and we add it to the SWE value, i.e. we update $x_t$ : we obtain the updated value $\tilde{x}_t, t = 1954, \ldots, 2014$ .

   (b) For each year, we perform ESP with historical climatology, i.e. we run the hydrological model.
   $\Rightarrow$ We obtain $\{M(\tilde{x}_t, \theta, I_t), t = 1954, \ldots, 2014\}$

   This two previous steps are repeated ten times leading to $61 \times 10$ predictions that we denote by $\{M^j(\tilde{x}_t, \theta, I_t), j = 1, \ldots, 10, t = 1954, \ldots, 2014\}$ At the end, we have a variability-corrected ESP forecast.

To assess the validity of the model, Talagrand histograms are used [5]. This histogram is obtained by :

1. Ordering the 10 predictions of the freshet volume, for each each year $t$.
2. Taking the percentile $q_t$ of the observed freshet volume $y_t$.
3. Repeating these two last steps for each year.
4. Drawing the histogram of $\{q_{1954}, \ldots, q_{2014}\}$.

The validity of the model is then assessed by analysing the histogram shape. Indeed, a U-shape shows that the model tends to under-estimate the true freshet volume because the true freshet volume more frequently correspond to the extreme percentiles. On the contrary, a "dome"-shape shows over-dispersion. Consequently, a efficient model results in a flat histogram, i.e. the distribution of the true freshet volumes $\{y_{1954}, \ldots, y_{2014}\}$ is uniform according to $\{M^j(\tilde{x}_t, \theta, I_t), j = 1, \ldots, 10, t = 1954, \ldots, 2014\}$.
The hypothesis that the distribution of the true freshet volumes $\{y_{1954}, \ldots, y_{2014}\}$ is uniform according to$\{M^j(\tilde{x}_t, \theta, I_t), j = 1, \ldots, 10, t = 1954, \ldots, 2014\}$ can also be tested by using the non-parametric test of Kolmogorov-Smirnov at the reference level of 5%. Indeed, if the p-value of the test is inferior to 5%, the hypothesis is rejected, otherwise it is not rejected.

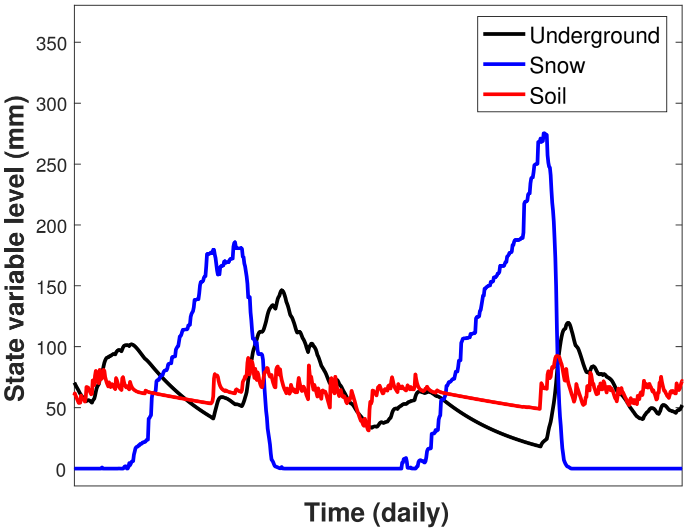


FIGURE 1 – Evolution of the state variable levels over two years.

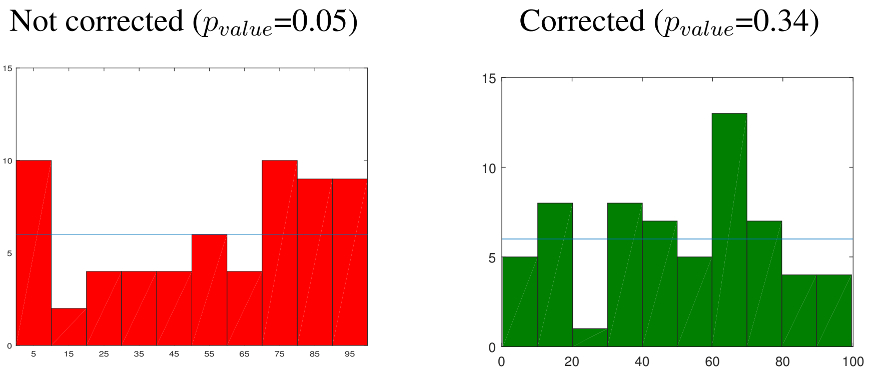
## 2.2 Extension to the summer

The aim is to apply this so-called $\Delta V$ method to the summer period. The first obvious obstacle is that the parameter with which one we play during the winter, the SWE, is of course not available in summer. Thus, the first step is to identify the one we will adopt.

The different state variables are the snow and the level of water on the underground and on the soil. As it can be seen in Figure 1, the snow has a behaviour which is pretty smooth. Thus, our idea is to pick a variable which has a quite similar behaviour. For this reason, we disregard the soil and decide to focus on the underground. Our first naive approach was to apply the previous method, the $\Delta V$ method, and so replacing SWE by the underground water level (UWL). Unfortunately, it is not very conclusive. If we compare the Talagrand histograms before and after applying the $\Delta V$ method, the histogram after correction does not seem to be flatter than the one before correction. Note that the Kolmogorov-Smirnov test does not reject the hypothesis of uniform distribution. However, it is well-known that the test can have insufficient power with small samples, i.e. its probability of rejecting the null hypothesis is low although this null hypothesis is false.

This outcome can be explained by the following reason : in winter, there is mainly only one phenomenon, the snow accumulation, while in summer, it is more complicated. Indeed, the evolution of the state variables actually depends on the level of water which is already present in the soil and underground. It is actually quite natural to understand. When it rains, if the area is very dry, it will not have the same consequences as if the area is already wet, which could lead to flood for instance.

The idea is then the following. In the **HINDCAST PART**, the two first steps are the same. Next, when building the DUWL distribution, we actually split the distribution into three distribu-

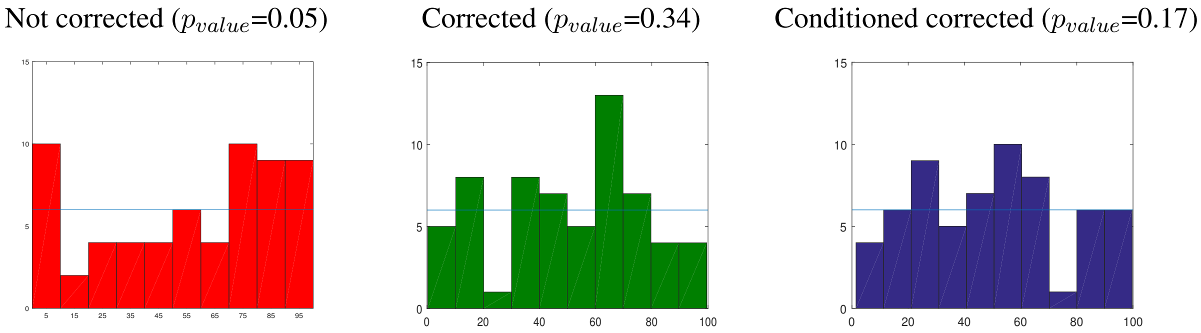


Not corrected ($p_{value}$=0.05)

Corrected ($p_{value}$=0.34)

FIGURE 2 – APPLICATION OF THE $\Delta V$ METHOD : COMPARISON OF THE MODELS BEFORE AND AFTER APPLYING THE $\Delta V$ METHOD

tion according to the state at the beginning (dry, medium or wet). The threshold to separate these situations are chosen by an empirical method based on the empirical distribution of data. After that, the **PREDICTION PART** is performed for each DUWL distribution.
This approach seems to give better results as it can be seen in Figure 3. Indeed, the Talagrand histogram showing the result after using both the $\Delta V$ method and taking into account the soil state is flatter.



Not corrected ($p_{value}$=0.05)

Corrected ($p_{value}$=0.34)

Conditioned corrected ($p_{value}$=0.17)

FIGURE 3 – Application of the $\Delta V$ method and conditioning : comparison of the models before and after applying the $\Delta V$ method and conditionning.

### 2.3 Discussion

Thus, this first approach, rather less naive than the first one, shows promising results that need to be investigated in more details.

## 3 Time Series Approach

Treating the hydrological model as a black box, we could choose to propose modifications to the inputs or the outputs of the model. Since the model output has to be adjusted by changing levels of state variables daily in order to match observations, we explored the idea that we might be able to model this adjustment using a time-series transfer function model.

ARMAX models are appropriate for predicting future values in a time-series as a function of past and present values of one or more exogenous variables. Let $Q_t^{sim}$ be the series of model predictions for a given time period and $Q_t^{obs}$ be the observed water levels for the same period.

We decided to focus on modeling the "error" series, $D_t$ as a function of initial conditions series and their lags, where

$$D_t = Q_t^{obs} - Q_t^{sim} \tag{3.1}$$

We initially chose to use the series of the sum over past 2 weeks of past precipitation values, $P_t$, as an exogenous variable in an ARMAX model. It is however possible to explore other available initial conditions series or to use several variables.

Such a model takes the form :

$$D_t = \frac{\nu(B)B^d}{\omega(B)} P_t + n_t \tag{3.2}$$

where $n_t$, the error term, is assumed to be a stationary time series of mean 0 that is uncorrelated with $P_t$ but might display an ARMA(p,q) autocorrelation pattern. The notation $B^d$ refers to the backshift operator that instructs us to choose the previous value in the series, ie.

$$Bx_t = x_{t-1} \;\; and \;\; B^d x_t = x_{t-d}$$

.

In this way, $\nu(B)$ and $\omega(B)$ are polynomial operators of finite order of the form :

$$\nu(B) = \nu_0 + \nu_1 B + ... + \nu_q B^q$$
$$\omega(B) = \omega_0 - \omega_1 B - ... - \omega_p B^p$$

When these polynomial operators are applied to the explanatory time-series $P_t$, they describe the lags and their associated coefficients that are relevant in modeling the response series.

The general form of the ARMAX model for the $D_t$ series can then be written in the perhaps more intuitive form :

$$\omega(B)D_t = \nu(B)B^d P_t + \omega(B)n_t \Longrightarrow D_t = \sum_{j=1}^{p} \omega_j D_{t-j} + \sum_{i=d}^{q+d} \nu_i P_{t-i} + \omega(B)n_t \tag{3.3}$$

for some finite lags p and q.

We can interpret this expression as saying that the difference between predicted water levels and observed water levels depends on past values of precipitation (or other initial conditions series) and past values of itself in a predictable way. This idea is philosophically aligned with the current practice of making adjustments to the initial conditions series each day in order to match the predicted and observed water levels. If we can estimate how the difference series ($D_t$) reacts to different values of initial conditions, we might be able to predict the discrepancy directly (instead of adjusting initial conditions to eliminate it ) and simply add this predicted difference to the model prediction to obtain a less biased and more adequately dispersed set of ensemble predictions.

For our model we chose $P_t$ as the exogenous variable and included lags of up to 7 days. The remaining noise model, $n_t$ was defined as having an ARMA(2,2) shape. In the future, other available variables can be included in the model and variable selection would have to be performed in order to be able to include larger lags, going all the way back to the previous years values possibly. Other available state variables and initial condition series are shown in Figure 4.

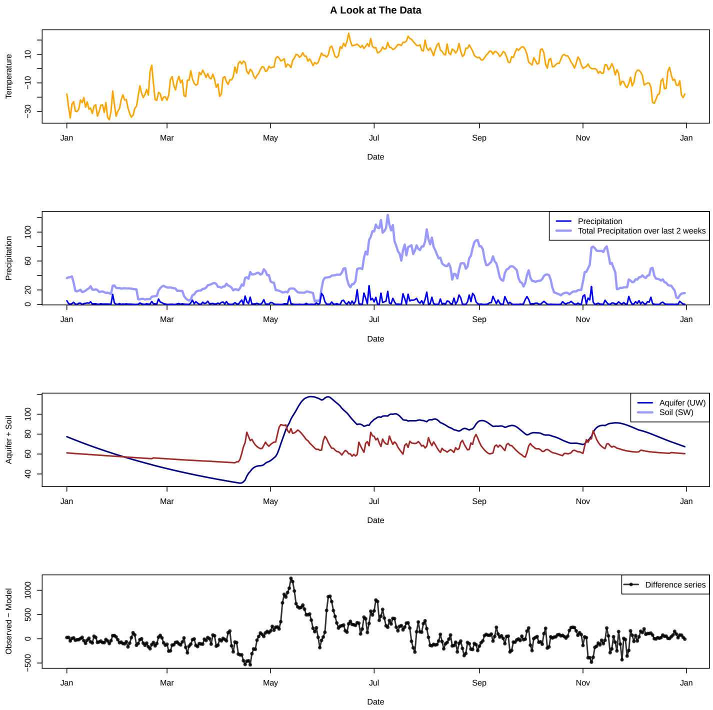


FIGURE 4 – Time-series of initial conditions and state variables (from year 1994) used by the CEQEAU hydro model to calculate inflow volumes in the three upper panels. These variables could be used as exogeneous variables in an ARMAX model to estimate the difference series ($D_t$), shown in the lower panel. We ultimately chose to base our model on the sum of precipitation values over past 2 weeks (from panel 2), that we refer to as $P_t$.

## 3.1 Simulation

For this project, we had access to data going back to 1954. For the estimation part of the model, we made use of weather and initial conditions time series ($T_t$, $P_t$, $SU_t$ and $SW_t$) , the output of the hydrological model before corrections ($Q_t^{sim}$) and observations of actual water levels for each year ($Q_t^{obs}$). Our prediction date was July 15th and we used data points occurring before this date to estimate model parameters for each year. We used the **armax** function of the **TSA R**[6] package to fit a transfer function model to the difference ($D_t$) series for each year, using the precipitation series, $P_t$ as the exogenous variable as in Equation 3.2. Parameters in the model were allowed to vary from year to year, though in future work some effort should be put into finding out if it is possible to come up with a general solution that can remain constant from year to year.

Once estimation was done, the model was transformed to a "regression" style model as in Equation 3.3. The reason for this is that the **armax** function lacks a **predict** method but the **arima** function from **stats** [7] package in **R** will accept a regression model with correlated errors and produce predictions. In this way, 63 different predictions for the $D_t$ series were made for the 15 days following July 15th, each one using the precipitation series for each year since 1954.

Recall that :

$$D_t = Q_{obs} - Q_{sim}$$

We then assume that the difference series is related to our predicted values as in :

$$D_t = D_t^{pred} + \epsilon_t$$

where $\epsilon_t$ that is assumed to follow a N(0, $V_\epsilon$) distribution. . We then retrieve a prediction for the expected water levels by adding the predicted $D_t^{pred}$ series for a particular year to the model prediction output for the respective year :

$$Q_t^{obs} = Q_t^{sim} + D_t^{pred} + \epsilon_t \Longrightarrow Q_t^{pred} = Q_T^{sim} + D_t^{pred}$$

In a subsequent step, we add 10 random white noise series to each prediction line to simulate the random noise, $\epsilon_t$, sampled from a normal distribution of mean 0 and variance $V_\epsilon$ estimated in the estimation step. In this way, we obtain 630 different scenarios that should have equal probability of occurring. Figure 5 shows the data used and the possibilities produced.

The total volume of inflows for the prediction period is then calculated for each scenario and compared to the observed inflow during this interval in order to produce the diagnostic PIT diagrams. Our results are illustrated in Figure 6.

## 3.2 Results and Discussion

The simple ARMAX model that we implemented during the workshop seems promising, as it improved the dispersion of the ensemble predictions as seen in Figure 6. However, there is much room for improvement.

First of all, most of the data present in the initial conditions time series would occur in the winter and spring (spring run-off is the dominant variable all the way up into May). If we wanted to characterize solely summer season data, we would be left with only a few data points for each year so we decided not to take that route.

The second problematic issue is related to model selection. We allowed the correlated noise series $n_t$ to follow an ARMA (p,q) model with p and q at most 2 while the exogenous variable $P_t$ could have an AR dependence of order 2 and MA terms up to order 7. Selection of variables was made in a heuristic manner, mainly by deciding what seems to work best among the options that are convenient to implement. However, in future work, a lot more focus should be put on variable selection. Since a lot of past data is available, we can imagine that lags of 1 year or more could play a role in model performance. Furthermore, since several initial conditions series are available, more than one exogenous variable could be used for fitting and predicting the difference series. It is computationally expensive to try to perform variable selection for such large sets of possible variables, therefore efficient algorithms for variable selection such as the adaptive LASSO would have to be investigated.

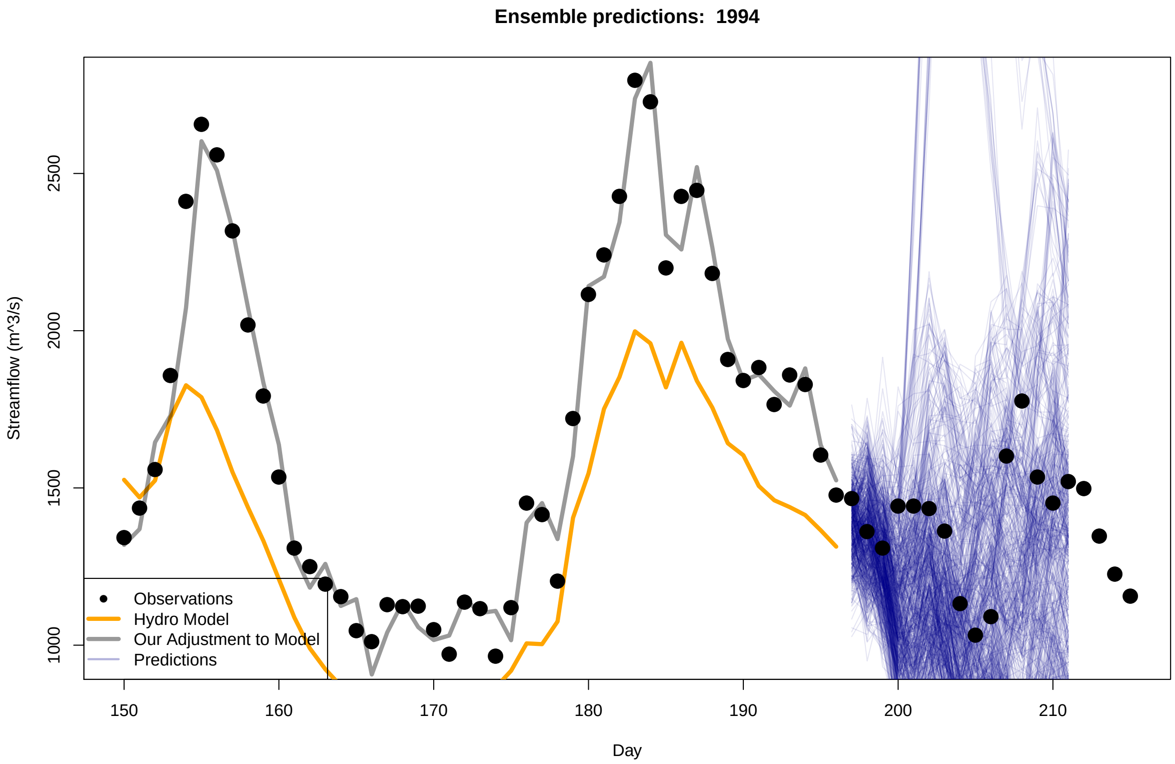


FIGURE 5 – Example of ensemble predictions for the year 1994 as output by hydrological model with our **armax** modifications. The blue lines each represent a predicted scenario based on weather data for one of the past 65 years with 10 different error series added on for a total of 650 shown possibilities.

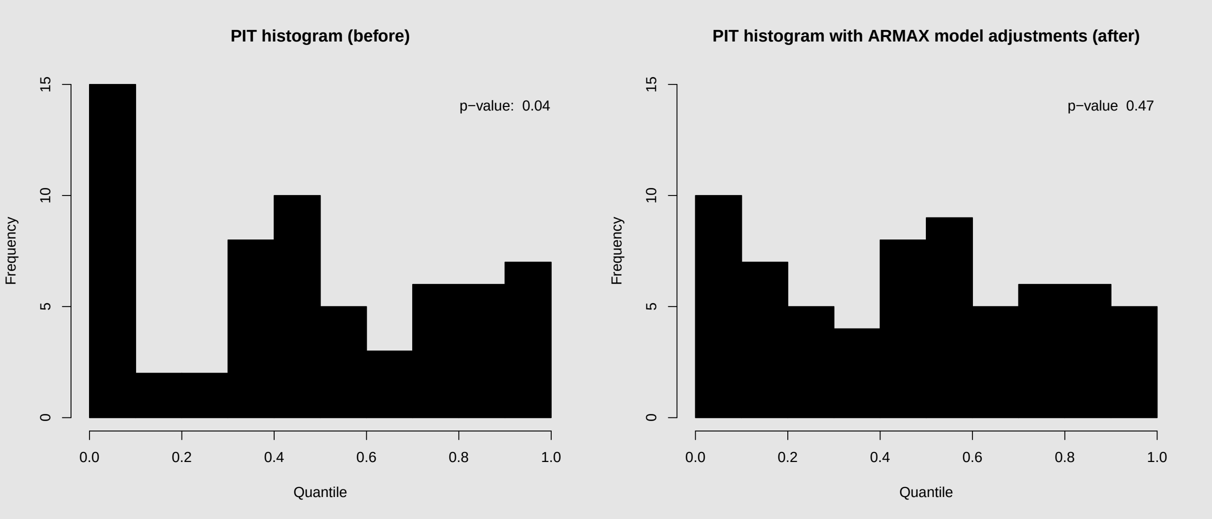


FIGURE 6 – PIT histograms for predictions before timeseries adjustment (left) and after (right). A Kolgomorov-Smirnoff test shows p = 0.04 for the original predictions and p = 0.47 predictions after the time-series adjustment was performed.

Another possibility that can be explored is related to the splitting of data. During the workshop, we decided to separate the data by year and fit data for each year separately. Two other possibilities deserve to be considered. The time series could be left "un-chopped" and we could attempt the fitting of a single model to all data points since 1954, allowing for long range correlations, with lags of up to several years. Another possibility is separating the data by year but fitting all years together in a multivariate time series model, with the purpose of obtaining parameters that are applicable to all years. This approach is not guaranteed to succeed (because it would need some underlying "physical" model to exist that connects the difference series to the initial conditions in a reliable way) but it deserves further investigation.

# 4 Gaussian Process Approach

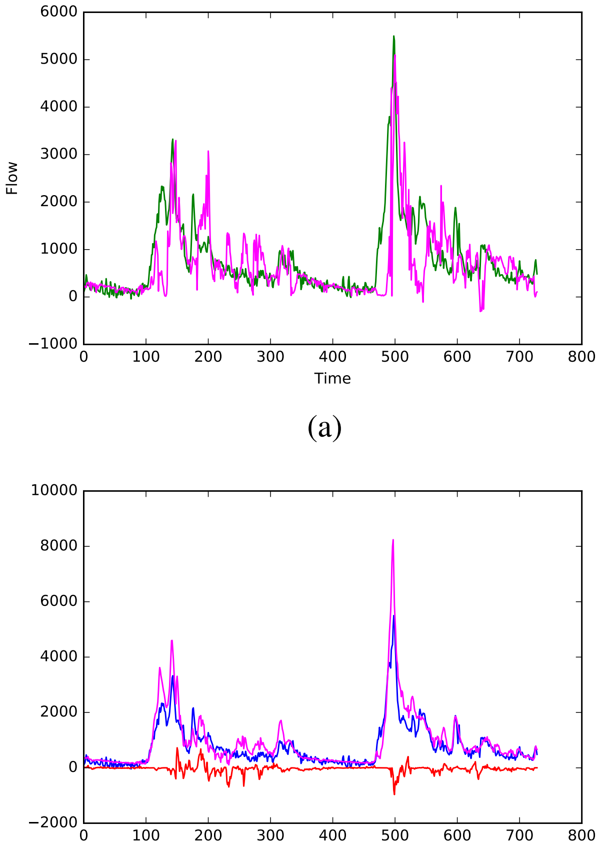
CEQUEAU is a complex model and its complex inner-workings are unclear. It is a deterministic code that takes as input the state variable and the weather information, denoted as $I$, and outputs the predicted flow $\hat{y}$. The Gaussian process is a popular tool for modelling the relationship between the input and the output of a complex computer simulation code [2]. This section demonstrates potential avenues where GP can be utilized for a subset of problems posed during the workshop. We begin with a brief background on GP (for a complete treatment on the topic, refer to [8]) followed by presentation of preliminary results. The report concludes with a discussion.

## 4.1 Gaussian Process

The first step is to view the complex computer code such as CEQUEAU as a black box function, $\hat{y}(x) = M(x)$, where $x$ denotes the input to the computer code and $\hat{y}$ denotes the output. The dependence between any pair of inputs, $x, x'$, are modelled using a covariance function, $K_\theta(x, x')$. The parameter $\theta$ captures the dependence between a pair of inputs $x, x'$ and it is to be estimated from the sample evaluations of the function : $(x_n, M(x_n))$, for $n = 1, ..., N$, where $N$ denotes the number of evaluations. For CEQUEAU model, each point $x_n = (I_n, t)$ is composed of the state variables and the weather forecast denoted by $I_n$ and $t$, the time of the year (for example, day of the year). There are various choices for the covariance function ; a popular choice is the radial basis function (RBF), also known as squared exponential kernel, which is suitable if the function $\hat{y}(x)$ is assumed to be a smooth function. The kernel function for squared exponential is given by,

$$K(x, x') = \exp\left(-\frac{1}{2\theta} d(x, x')\right),$$

where $d(x, x')$ denotes the distance between the two inputs $x, x'$. Typically, $d(x, x') = \|x - x'\|^2$. Since there is little reason to assume that CEQUEAU model would vary widely for a small change in the input, the squared exponential kernel was used in carrying out the preliminary results.



(a)

(b)

FIGURE 7 – Prediction results on the unseen data. (a) Green indicates the calibrated values outputted by CEQUEAU and magenta indicates the predicted values outputted by GP. (b) Blue indicates the observed flows, magenta denotes the output of CEQUEAU model plus the predicted residual values outputted by GP, and the red indicates the residuals outputted by GP.

## 4.2 Results

The points at which to evaluate the black box function are typically determined via carefully determined statistical design [9]. However, technical issues prevented direct evaluation of the CEQUEAU model during the workshop. However, we were provided access to calibrated output from CEQUEAU model and the preliminary results were obtained by fitting GP to this data. Additional exploration consisted of fitting GP on the residuals obtained between the calibrated output of the CEQUEAU model and the observed flow.

GP was fitted to the calibrated CEQUEAU model on the 10 years data starting from 1953-Jan-01 to 1962-Dec-31. The estimated parameter is $\theta = 0.405$. This fitted model is used to predict on the 2 years data starting from 1963-Jan-01 and the result is shown in Figure 7 (a). The prediction seems to model the output of CEQUEAU quite well in its ability to pick out the peaks and troughs. To improve upon this, we have also fitted the GP on the residual, $r(x_t) = y_t - \hat{y}(x_t)$, where $y$ denotes the observed flow at time point $t$. The prediction on the unseen data from

1963-Jan-01 for 2 years is shown in Figure 7 (b). This figure seems to indicate that modeling the residual leads to accurate modeling of the observed flows. It illustrates that perhaps the best use of GP is in combination with CEQUEAU. That is, rather than modeling for the output of CEQUEAU directly but rather, model for the difference between the observed and the calibrated values outputted by CEQUEAU. However, this conclusion may be premature considering that we did not have access to the CEQUEAU code.

## 4.3 Discussion

We will conclude this section with a discussion on solving one of the main difficulties underlying the use of CEQUEAU model : determining the suitable initial condition, $z_0$, where $z$ denotes the underlying state of the earth (i.e., $I = (z, w)$ where $w$ denotes the weather data). The current procedure encapsulates manual trial-and-error type approach where a technician tries multiple values of $z_0$ until the simulated flow appears close enough to the observed flow ; for example, $\sum_{t=1}^{T} \|y_t - \hat{y}_t\|$ is *small*, where $t = 1, ..., T$ denotes the total number of time points under consideration. The precise definition of small is unclear and the procedure relies heavily on the experience of the technician and his prior knowledge of how the CEQUEAU model works. GP may provide a direction in inferring the initial condition, $z_0$.

The GP model can be useful as a tool for simulation. Starting from an arbitrary initial value, we can use GP to simulate the forecast for the next 14 days. And it may be possible to simulate the future values using GP for optimization of the initial state. That is, we can pose the problem of finding the suitable initial value as an optimization problem :

$$\hat{z}_0 = \arg\min_{x_0} \sum_{t=1}^{T} (y_t - \hat{y}(z_t | z_{t-1}, w_{t-1}))^2; \tag{4.1}$$

where $\hat{y}(z_t | z_{t-1}, w_{t-1})$ is simulated by the GP model. The solution to the above optimization problem may be obtained via Bayesian optimization techniques [3]. This optimization can be performed over the collection of past data. Also note that posing the problem of finding the initial condition as an optimization problem makes clear the notion of optimality in terms of the choice of the initial values (i.e., in minimizing the squared difference between the observed and the predicted values).

Note also that the state variables $z_t$ are hidden variables, that evolve according to hidden Markov model structure :

$$z_t = f(z_{t-1}, w_{t-1}) + \eta_t, \text{ for } t > 0, \tag{4.2}$$

where $f$ is a function of the current state $z_{t-1}$ and the weather condition $w_{t-1}$ and $\eta_t$ denotes the random fluctuation. The observed values are function of the state variables :

$$y(z_t) = g(z_t) + \epsilon_t, \tag{4.3}$$

where $\epsilon_t$ denotes the random noise. A sequential Monte Carlo methods [4] are widely adopted for problems exhibiting hidden Markov model structure. Formulating the problem of finding the initial condition in the context of SMC may be the next step towards formalizing this problem of inferring the initial value.

# 5 One dimensional model

The idea behind the development of a one-dimensional model is to produce a simple, but realistic, deterministic hydrological model which will allow us to develop an understanding of the water flux and which can then be used to test various data assimilation approaches.

## 5.1 The model fomulation

In the one-dimensional model we consider the whole hydrological basin to be a series of connected *cells* $C_i$. In this model $C_1$ will be assumed to be the cell which is highest in altitude, and that there is a positive free water flux $F_i^n$ from cell $C_i$ to cell $C_{i+1}$, with no flux into cell $C_1$. The cell $C_i$ will contain a quantity $SL_i^n$ of ground water in the soil, and $SW_i^n$ of snow, at the time $t_n$. In addition, each cell will be subjected to a temperature $T_i^n$ and a precipitation $R_i^n$ which can be obtained from measured data. This is illustrated in Figure 8.

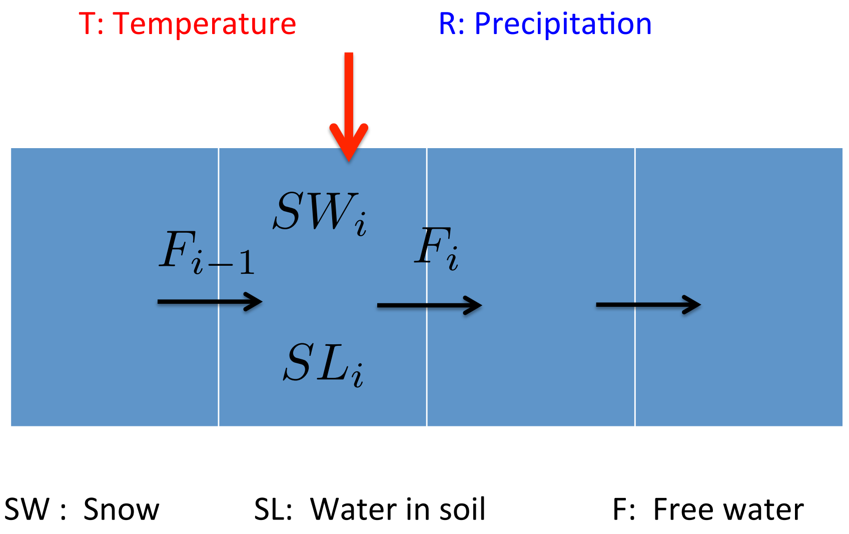


FIGURE 8 – A schematic of the one-dimensional cell model

We now make the following physical assumptions

1. The free water flux is a function of the water in the soil. So that

$$F_i^n = f\left(SL_i^n\right). \tag{5.1}$$

2. If $T_i^n > 0$ then the snow melts as the temperature rises and the melt water adds to the water $SL_i^n$ in the soil. In addition the precipitation falls as rain and adds to the soil water.
3. If $T_i^n \leq 0$ then the precipitation falls as snow and adds to the snow in the cell $SW_i^n$.

We make the following assumptions

— The amount of snow that melts in one time unit is directly proportional to the temperature above freezing. All of the melted snow becomes soil water.

— If the temperature is *positive* then the amount of soil water added in a time unit is directly proportional to the precipitation.

— If the temperature is *negative* then the increase in the snow in the cell in one time unit is directly proportional to the precipitation.

The resulting mathematical model is then as follows

If $\quad T_i^n > 0 \quad$ then

$$SW_i^{n+1} = SW_i^n - \alpha T_i^n \, SW_i^n \tag{5.2}$$

$$SL_i^{n+1} = SL_i^n + \alpha T_i^n \, SW_i^n + \beta R_i^n + F_{i-1}^n - F_i^n. \tag{5.3}$$

If $\quad T_i^n \leq 0 \quad$ then

$$SW_i^{n+1} = SW_i^n + \gamma R_i^n \tag{5.4}$$

$$SL_i^{n+1} = SL_i^n + F_{i-1}^n - F_i^n. \tag{5.5}$$

We will assume that there are $N$ such cells. As described above, we take $F_0^n = 0$. The value

$$Q^n = F_N^n$$

is then the *measured flux* which will be our primary measurement for the data assimilation calculation.

## 5.2 Implementation

To implement this model, we take a time unit of *one day*, and take $N = 10$ cells. The values of the constants of proportionality above, and the function in (5.1) were chosen to get a reasonable fit to the measured data and we took

$$\alpha = 1/15, \quad \beta = \gamma = 1, \quad f(S) = S/4. \tag{5.6}$$

The temperature and precipitation data were taken from those supplied. Initial values of $SW$ and $SL$ were estimated and the model was 'spun up' by running it over several years before results were plotted. The resulting model was very easy to implement in Matlab. As output we took $SW^n$ to be the *total* amount of snow in all of the cells combined on the nth day, and we also calculated $Q^n$. The results of these calculations in one year, together with the mean temperature $T^n$ and mean precipitation $R^n$ are presented in Figure 9. Similarly, calculations over ten years are presented in Figure 10.

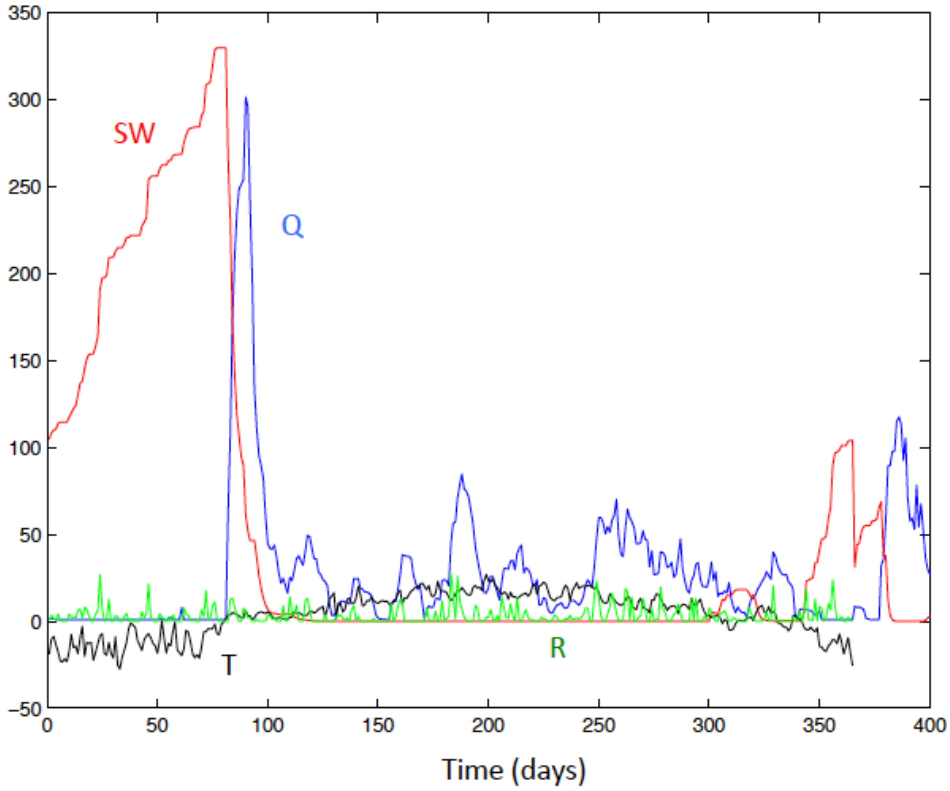


FIGURE 9 – The output of one year of the one-dimensional cell model, plotting outputs $Q$ (blue) and $SW$ (red) and inputs $T$ (black) and $R$ (green).

The results of a one year simulation shown in Figure 9 show that the total $Q$ has a sharp peak in the Spring when the snow melts, and that this is the dominant effect on its value through the year. The (rather unpredictable) rainfall $R$ has a lesser, though still significant, effect. A similar result can be seen in the ten year simulation, and we see a variation in the total amount of snow $SW$ from one year to the next.

## 5.3 Data assimilation

Noting from the above calculation that the total amount of snow fall $SW$ has the main effect on the total flux $Q$, we ask the question of how well is it possible to estimate $SW$ from the data ? Accordingly, we conducted the following test which addresses the question of how accurately we can find an initial value of $SW$ from noisy measured data of the total flux $Q$.

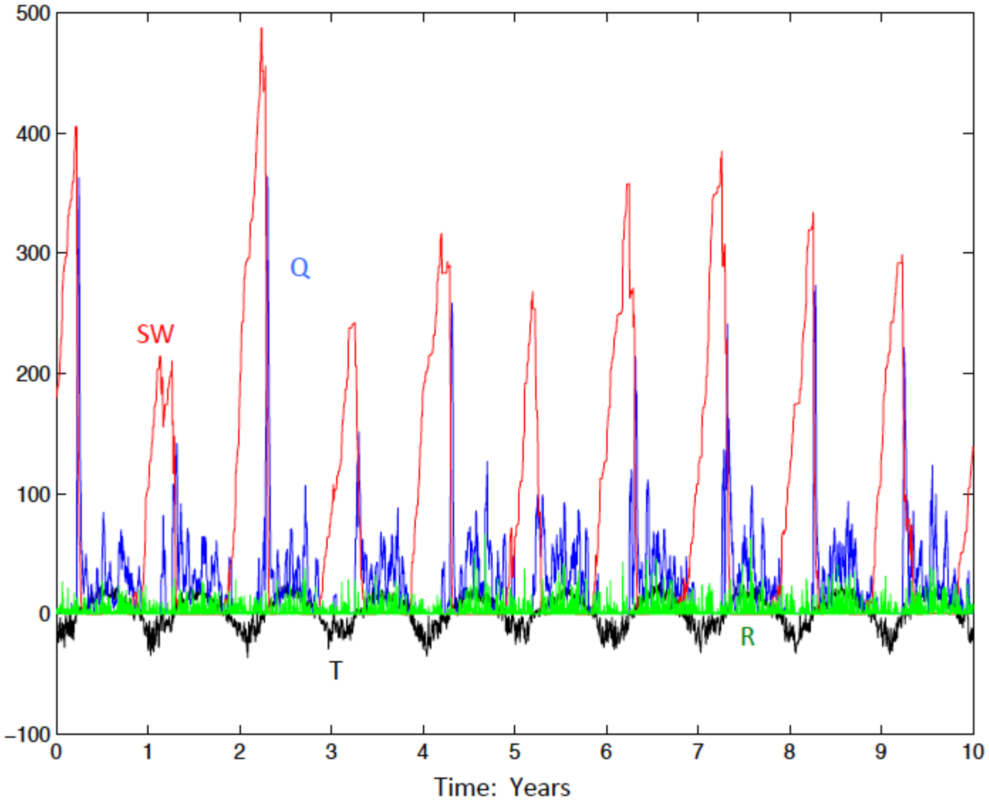


FIGURE 10 – The output of ten years of the one-dimensional cell model.

- Take an initial *estimated value* of $SW^0$ at time $t^0$ (assumed to be in Winter) and distribute the snow evenly over all of the cells. Set the initial value of $SL = 0$ (this is reasonable for winter months).
- Run the model to generate two years of values for the total flux $Q^n$.
- Add noise to $Q$ to give the noisy measured data $Q_{Noise}$.
- Take a shift $SW^0 \to SW^0 + \delta SW$ and generate a new time history $Q_\delta$ of the total flux.
- Find $\delta S$ which minimises the *error* between $Q_{Noise}$ and $Q_\delta$.

An example of the difference between $Q$ and $_{noise}$ is given in Figure 11.

We considered two measures of the error

$$E_1 = \left( \int (Q_{Noise} - Q_\delta) \; dt \right)^2 \tag{5.7}$$

and

$$E_2 = \int (Q_{Noise} - Q_\delta)^2 \; dt. \tag{5.8}$$

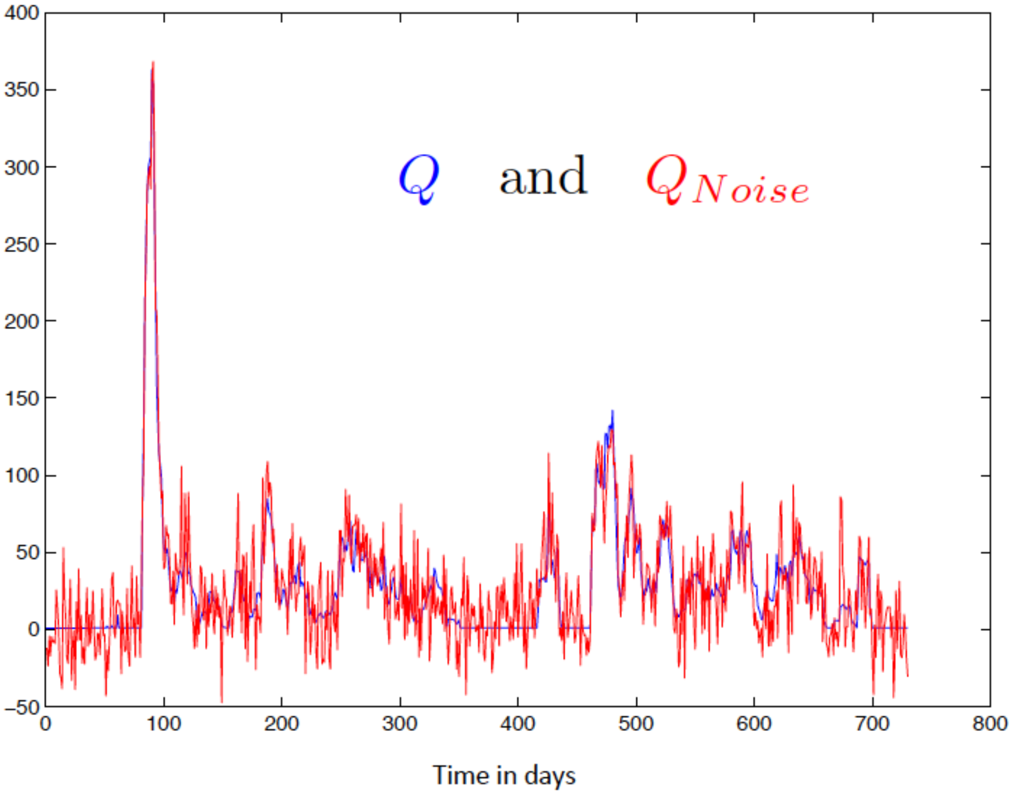


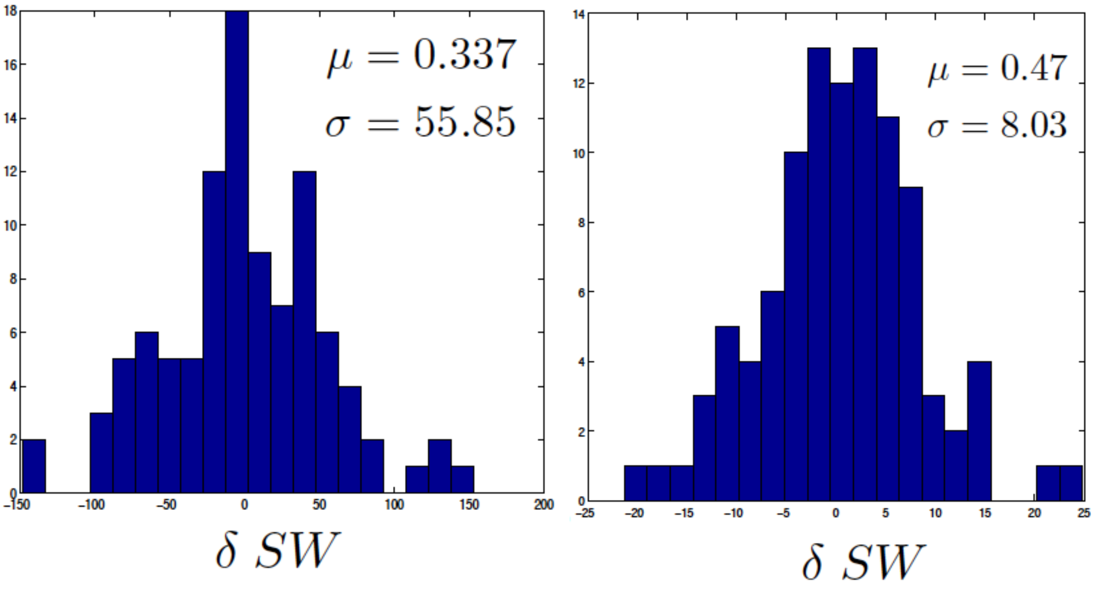
FIGURE 11 – The output $Q$ of two years of the one-dimensional cell model, together with a noisy perturbation $Q_{Noise}$

In the case of $E_1$ we choose $\delta S$ so that $E_1 = 0$. In the case of $E_2$ we choose $\delta S$ to minimise its value.

*We ask the question of which of these two error measures leads to the better estimate of $SW$.*

We can test this procedure by considereing a number of realisations of $Q_{Noise}$ with random noisy data. If the data assimilation procedure is working well, then over all of these realisations we should see a mean of $\delta S = 0$. A measure of the performance of this algorithm is given by the standard deviation $\sigma$ of the resulting estimate. Accordingly we take 100 realizations, with a Gaussian noise added to $Q$.

The resulting histogram of the values of $\delta SW$ is shown in Figure 12 together with estimates of the mean $\mu$ and standard deviation $\sigma$. We show two plots with Figure 12 (a) the estimate for error measure $E_1$ and Figure 12 the estimate for error measure $E_2$. We can see that whilst the



(a) E1

(b) E2

FIGURE 12 – The histogram of the estimates of $\delta SW$ given by the two error measures $E_1$ and $E_2$. The mean and standard deviation are also given. We see that the error measure $E_2$ gives a much better estimate of $SW$ than $E_1$

mean of the two estimates is close to zero in both cases, that in contrast the standard deviation of the error for measure $E_1$ is *much* higher than for measure $E_2$. We conclude that in the data assimilation routine the measure $E_2$ gives a much better estimate for the initial snow value $SW^0$.

## 5.4 Discussion

Whilst very simple, the one-dimensional model gives surprisingly realistic looking results for the variation in the amount of snow and the total flux. It is also useful in testing the two error measures used in the data assimilation calculation to estimate the initial snow value from the measured flux $Q$. Further tests of the simple model could be to see how well the future values of the flux (which are what we are interested in) can be predicted. I recommend this model be used to estimate the effectiveness of more data assimilation procedures. Its simplicity allows it to be used for a large number of realizations and to test many schemes, in a manner that may not be possible with the more complex CEQUEAU model. It would also be interesting to do a comparison between the predictions of the one-dimensional model and these full simulations.

# 6 Acknowledgements

We would like to thank Richard Arsenault and Marco Latraverse from Rio Tinto for bringing us this interesting problem, Université de Montréal for the delicious lunches, Dr. Odile Marcotte for organizing the workshop and Prof. Pierre Duchesne for co-ordinating our team. We would also like to thank Diana Jovmir for taking the main responsibility of the writing of this report.

# Références

[1] Richard Arsenault, Marco Latraverse, and Thierry Duchesne. An efficient method to correct under-dispersion in ensemble streamflow prediction of inflow volumes for reservoir optimization. *Water Resources Management*, 30(12) :4363–4380, Sep 2016.

[2] Derek Bingham, Pritam Ranjan, and William J Welch. Design of computer experiments for optimization, estimation of function contours, and related objectives. *Statistics in Action : A Canadian Outlook*, 109, 2014.

[3] Eric Brochu, Vlad M Cora, and Nando De Freitas. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv :1012.2599*, 2010.

[4] Arnaud Doucet and Adam M Johansen. A tutorial on particle filtering and smoothing : Fifteen years later. *Handbook of nonlinear filtering*, 12(656-704) :3, 2009.

[5] TM Hamill. Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129(1) :550–560, 2001.

[6] R Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[7] R Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[8] Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*, volume 1. MIT press Cambridge, 2006.

[9] Jerome Sacks, William J Welch, Toby J Mitchell, and Henry P Wynn. Design and analysis of computer experiments. *Statistical science*, pages 409–423, 1989.