Commercial greenhouse water demand sensitivity analysis: single crop case study

Dean C. J. Rice, Rupp Carriveau and David S.-K. Ting

ABSTRACT

Today water distribution utilities are trying to improve operational efficiency through increased demand intelligence from their largest customers. Moving to prognostic operations allows utilities to optimally schedule and scale resources to meet demand more reliably and economically. Commercial greenhouses are large water consumers. In order to produce effective forecasting models for greenhouse water demand, the factors that drive demand must be enumerated and prioritized. In this study greenhouse water demand was modeled using artificial neural networks trained with a dataset containing eight input factors for a commercial greenhouse growing bell peppers. The dataset contained water usage, climatic and temporal data for the years 2012–2014. This model was then evaluated using the Extended Fourier Amplitude Sensitivity Test, a global sensitivity analysis, in order to determine the importance, or sensitivity, of each input factor. It was found that time of day, solar radiation, and outdoor temperature (°C) had the largest effects on the model output. These outputs could be used to contribute to the generation of a simplified demand-forecasting model.

Key words | commercial greenhouse watering, neural network, sensitivity analysis, water demand characterization

NOTATION

The following symbols are used in this paper:

- **ANN**: Artificial neural network
- **$A_j$**: Fourier coefficient
- **$B_j$**: Fourier coefficient
- **$D_{\text{FAST}}$**: Total output variance of the model
- **$\tilde{D}_{-i}$**: Variance caused by all factors except $i$th input factor
- **$\tilde{D}_{ii}$**: Variance caused by $i$th input factor
- **$F_i$**: Forecast or modeled value of $i$th parameter
- **GSA**: Global sensitivity analysis
- **$L$**: Litre
- **LSA**: Local sensitivity analysis
- **$M$**: Interference factor, usually taken as 4
- **$N$**: Number of data points
- **$O_i$**: Observed value of $i$th parameter
- **SCADA**: Supervisory control and data acquisition
- **$S_i$**: First order sensitivity indices
- **$S_{ST}$**: Total order sensitivity indices
- **$X$**: Input factor
- **$Y$**: Model output
- **ha**: Hectare (10,000 m$^2$)
- **$r$**: Pearson product-moment correlation coefficient
- **$s$**: Activation variable for eFAST transformation, taken uniformly between $\pm \pi$
- **$x_i$**: Transformed value of input $i$
- **$\phi$**: Random phase-shift used in eFAST input transformation
- **$\omega_i$**: Incommensurate frequency assigned to input $i$

INTRODUCTION

The primary mandate of water utilities is to provide a safe, uninterrupted supply of potable water. This can often make network optimization a lower priority. However, water
utilities can consume up to 40% of a municipality’s energy bill, of which approximately 80% is used for distribution (pumping) (United States Environmental Protection Agency 2015). Consequently, it is clear that inefficient pumping schedules can be costly. The development of demand forecasting models plays a key role in operational habits and can be vital in developing pumping schedules that allow the network to operate at equilibrium (Fodya & Harley 2014). Pumping schedules are not the only areas which demand forecasting can impact; network upgrades can be very expensive, consequently, it should be certain that these upgrades are critically necessary. Network upgrades are often proposed based on network hydraulic models (Jain & Ormsbee 2002), which can identify areas of interest such as low-pressure zones. There can be some vulnerability with this approach as the hydraulic model is based on estimates of demand, estimates that are typically broadly based on low-resolution billing figures. This relative inaccuracy can then be propagated forward, leading to less than optimum network upgrade forecasts.

Numerous studies have been undertaken to provide more accurate methods for water demand forecasting, Donkor et al. (2014) have provided a comprehensive review of many of these studies which took place between the years 2000 and 2010. The study outlined three main concerns that were not addressed in the works studied:

- **Model practicality**: Is the model easily implemented by the utility?
- **Input selection**: Are the proposed model inputs/indicators easily/inexpensively monitored?
- **Input importance**: Are all proposed inputs/indicators necessary and have they been prioritized?

The necessity of model practicality cannot be overemphasized. Most utilities are streamlined facilities that are often challenged by regular day-to-day operation and do not have the resources to manage complex operational models. The selection of easily monitored input factors is also critical. Some inputs may be deemed relevant, but, if they cannot be measured well, or easily, it limits the model’s practicality. This is particularly true of models with high temporal resolution. Examples of such factors range from gross national product and inflation rate (Firat et al. 2009), appliance ownership and efficiency (Williamson et al. 2002), and household size and income (Polebitski & Palmer 2010).

The significance of input factor priority, as stated by Donkor et al. (2014), is to ensure inputs are chosen that have notable influence on the model output and do not create spurious relationships. There is, however, another purpose for the selection or screening of input variables, and that is computational cost. The inclusion of non-essential input variables in a model can be computationally expensive as the model attempts to determine relationships that may be very weak or nonexistent between extraneous variables. Such an expense can translate into an increase in the computational time and may also compromise accuracy through overgeneralization of the model.

A major impetus for the study was to address the concerns of Donkor et al. (2014) while trying to characterize the major influencing factors for large water consumers in greenhouse agriculture. Figure 1 shows the capacity breakdown by consumer type for a large water utility in Essex County Ontario. Inspection of the figure reveals that over three-quarters of all water goes to commercial greenhouses. It is not difficult to imagine how this region of southwestern Ontario has the densest concentration of commercial greenhouses in North America.

The planned approach is to model greenhouse water usage behaviour using an artificial neural network (ANN) and screening input factors using global sensitivity analysis (GSA). GSA will allow the underlying relationships between

---

**Figure 1** | Capacity breakdown by consumer type for water utility (Other: Industrial, commercial excluding greenhouses).
inputs and outputs of the model to be quantified making it possible to remove input factors that have little to no effect on the output. This technique will provide a basis for model simplification and allow the model developer to focus on using inputs that are easily monitored by the water utility without compromising model accuracy. For this water utility, providing an easily executable and more accurate demand forecasting model for greenhouse water usage will greatly improve day-to-day operations, enable optimization of pumping schedules, and promote improved infrastructure planning. This procedure is not unique to greenhouse water demand modeling; it has great potential in water distribution networks that are dominated by a single sector, e.g. lumber/wood products, petroleum, and oil refinement.

Data screening

Data screening is a term used to describe the process of filtering out inputs that have little to no effect on the output of a model (Morris 1991; Saltelli et al. 2004). Fu et al. (2012) explored the use of GSA, more specifically Sobol’s method, on two water distribution networks to ensure only variables which had significant effects on the output were included in the model. Sobol’s method is a variance-based method for quantifying the effect each input has on the output while also taking into account the interactions between input variables. It was found that removal of these insensitive factors allowed the model to become more computationally efficient while not compromising accuracy. This study will use the Extended Fourier Amplitude Sensitivity Test (eFAST) as the method of GSA. The choice of eFAST was based on recommendations of Saltelli et al. (1999) for its lower computational cost and similar accuracy when compared to the method of Sobol, and eFAST has been used in studies pertaining to wind turbine power output (McKay et al. 2014), crop growth models (Wang et al. 2013; Vanuytrecht et al. 2014) and water treatment models (Cosenza et al. 2013). Further details on GSA will be provided later in this paper.

Study area

This study will examine a service area of approximately 90,000 ha located in Essex County, Ontario, Canada (Figure 2). Ontario is the home of 830 ha of commercial vegetable greenhouses, of which 630 ha are located in Essex County, making it the area with the highest concentration of greenhouses in
North America (OGVG 2014). This poses a unique challenge for the water utility since their distribution system is dominated by one industry. This challenge is that any advancement in technology or change in process in the industry can cause a serious imbalance in the distribution system. This is why it is important to characterize water demand for large consumers and to re-evaluate these models periodically.

**Greenhouse water usage studies**

There have been several studies conducted on agricultural water needs; however, research on commercial greenhouse water usage is extremely limited. This section will discuss studies that have been undertaken that involve crop water usage and may not incorporate greenhouses. Orgaz et al. (2005) examined plant water demand in unheated plastic greenhouses in order to determine crop coefficients to be used to calculate evapotranspiration rates, which is the sum of soil water evaporation and plant transpiration. The procedure was carried out for four prevalent local crops (melon, watermelon, sweet pepper, and green bean) with a soil growth medium. It was shown that there were considerable differences in water requirements dependent on crop growth stage, season, and growing practices. In a United Nations report (Doorenbos & Pruitt 1977) it was found that solar radiation had the largest effect on evapotranspiration rates of various crops. Ma et al. (2015) also examined the environmental factors influencing water evaporation, soil water evaporation, and plant transpiration. Using regression analysis, equations were developed to describe the behaviour of these evaporative processes based on three climactic factors: indoor temperature, indoor humidity, and solar radiation. A genetic algorithm was then used to optimize these equations. In this case the genetic algorithm was used to determine the minimum values of evaporation and transpiration and the values of the corresponding climactic factors at which they occur. These idealized climatic factors could then be implemented into the artificial environment of the greenhouse in order to reduce plant-watering needs. The main focus of these studies along with several others (Thompson et al. 2007; Capraro et al. 2008; Bernier et al. 2010) was to improve watering schemes inside the greenhouse operation. These studies are beneficial to the greenhouse operations in terms of water conservation, but do not address the needs of the water utility, which require a more general view of greenhouse water usage in order to forecast water needs.

**Greenhouse operations**

In this region, greenhouse operations use large storage tanks to supply the crop with water. These tanks are typically filled in the evening when demands on the water network are low. This raises an issue when modeling greenhouse water demand since the water usage inside the greenhouse will not be reflected in consumption from the water utility until the storage tanks are refilled. Another facet of greenhouse watering is recycling, which is the process of collecting and treating runoff or excess water and mixing it with fresh water in order for it to be used again in the crop watering process. This process allows for the efficient use of fresh water and reduces the associated costs. The limit to the reuse of this excess water in most cases is salinity levels, more specifically sodium chloride (NaCl) (Trajkova et al. 2006). The salinity tolerances vary from crop to crop and in order to ensure these values are not exceeded, salinity levels are measured in the greenhouse using the electrical conductivity of the water. Greenhouse operation can also utilize alternative water sources such as wells and ponds. Usage of these alternative sources can dramatically impact the amount of water needed from the utility, but can also pose issues to crop health. Use of untreated water such as recycled, rain, and pond water can potentially destroy crops, as there is no way of knowing what contaminants it contains, which could cost the grower millions of dollars. Because of this risk, most operations in this region employ the use of small-scale water treatment facilities. These treatment facilities are very limited in capacity as the costs associated with larger-scale options outweigh the costs of potable municipal water; this combined with evolving regulations on alternative water sourcing solidifies the need for municipal water sources in greenhouse operations. This study examines the use of water inside the greenhouse, meaning the water that has been sent to the plants. The methods and technologies mentioned above are not used in every greenhouse operation worldwide so examination of plant watering trends will be of use to any water utility dealing with greenhouse demand.
THE DATA

The analysis utilizes data reported every 15 minutes for each of the factors in Table 1. The data are for 1.42 ha of greenhouse growing bell peppers and cover the years 2012, 2013 and 2014. The data were collected from the supervisory control and data acquisition system (SCADA) of the greenhouse operator and contain 100,609 data points for each factor. The greenhouse logs numerous factors such as water electrical conductivity, fertilizer levels, and growth medium weight along with climactic and temporal data. It should be noted that the data were collected from a heated greenhouse, meaning the greenhouse is heated to a minimum of 20°C during the winter months with a few exceptions occurring during the cleanout process at the beginning and end of the year. It is obvious that many of these factors require instrumentation inside the greenhouse operation, which would be difficult if not impossible for the water utility to implement on a large scale. The dataset does however contain two factors that would require internal (inside the greenhouse) monitoring; these factors are greenhouse humidity and greenhouse temperature. The purpose of using these internal factors is to compare their importance to that of the external (measured outside the greenhouse) factors and to determine if the water demand can be reliably forecast using only external factors.

Table 1 | Input factors used in MATLAB neural network

<table>
<thead>
<tr>
<th>Input factor</th>
<th>Range (Min–Max)</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greenhouse Temperature</td>
<td>2.11–37.11</td>
<td>Celsius (°C)</td>
</tr>
<tr>
<td>Outdoor Temperature</td>
<td>−23.30 to 34.51</td>
<td>Celsius (°C)</td>
</tr>
<tr>
<td>Cumulative Solar Radiation</td>
<td>0–3,096</td>
<td>Joule per square centimetre (J/cm²)</td>
</tr>
<tr>
<td>Solar Radiation</td>
<td>0–1,045.65</td>
<td>Watt per square metre (W/m²)</td>
</tr>
<tr>
<td>Wind Speed</td>
<td>0–13.69</td>
<td>Metre per second (m/s)</td>
</tr>
<tr>
<td>Greenhouse Relative Humidity</td>
<td>31–100</td>
<td>Percent (%)</td>
</tr>
<tr>
<td>Time</td>
<td>0–23.75</td>
<td>Decimal hour</td>
</tr>
<tr>
<td>Month</td>
<td>1–12.97</td>
<td>Decimal month (Jan 01 = 1, Dec 31 = 12.97)</td>
</tr>
</tbody>
</table>

The rationale for the use of external greenhouse factors is that a water utility can easily monitor them with the installation of a small-scale weather station at their distribution centre. One issue that may arise is the double layer of polyethylene used in greenhouse construction. This material will diffuse the solar radiation leading to different values being recorded by indoor and outdoor sensors. For this analysis, the numerical value of the solar radiation is of little importance, as this study will examine the effect input factor variation will have on the output of the model (water usage).

The entire data collection system is connected to the greenhouse operator’s commercial SCADA system and all data are collected in one software package. Greenhouse operations are divided into zones and each zone has its own water feed. Each feed contains metering measuring flow, temperature and various other characteristics mentioned previously, which are fed into the control software for analysis. The water usage data used in this analysis are measured as cumulative water usage in litres every 15 minutes with flow sensors, for which the data reset every day at 700 hours.

Preliminary data analysis

Before performing a sensitivity analysis (SA), it is crucial to carry out more qualitative methods of data analysis in order to obtain a basic sense of the relationships between input and output factors. Figure 3 shows the monthly average and maximum 15-minute water usage over the entire dataset. It can be observed that water usage increases from January until it peaks between June and July, after which the water usage decreases until the end of the year. This pattern can be explained with insight into the operational habits of the greenhouse. In January, new crops are installed in the greenhouse, which consume less water than fully grown plants. The peak of the pepper growth cycle occurs during July and August when the largest water consumption is observed. Pepper growth then declines from September through November when the plant life cycle is at an end. During the month of December there is no pepper production, but water is used during the process of removing old plants and cleaning out the greenhouse to prepare for the future growing season. Further examination of Figure 3 shows the large spread between the average and maximum 15-minute water usage. The reason for this is that watering...
is not constant; there are a high number of occurrences of zero water usage over the 15-minute intervals, which bring the average down drastically.

Table 2 summarizes the correlation coefficients for each factor. It can be observed that a value of 1 exists along the diagonal, which shows that the correlation between the factor and itself is perfectly linear. Table 2 will also provide a basis for validation of the sensitivities that will be produced using the GSA. The results in Table 2 show that the input factor that has the strongest linear relationship with water usage (output) is solar radiation ($r = 0.753$). Weak correlation exists between water usage and greenhouse temperature ($r = 0.453$) and humidity ($r = -0.404$). Table 2 also exposes relationships between input factors. Greenhouse temperature and solar radiation have the strongest linear relationship ($r = 0.617$). Greenhouse temperature and humidity ($r = 0.547$) and solar radiation and humidity ($r = 0.534$) show moderate correlation. The existence of these moderate-to-strong correlation coefficients between input factors show that there is multicollinearity within the inputs. Multicollinearity can cause issues when modeling, particularly in linear regression models (De Veaux & Ungar 1994), as it can increase the variance of coefficient estimates and makes these estimates sensitive to minor changes. How these effects are to be dealt with will be addressed in a later section.

**Table 2 | Pearson product-moment correlation coefficients for all factors**

<table>
<thead>
<tr>
<th></th>
<th>Month</th>
<th>Time</th>
<th>Humidity</th>
<th>Solar radiation</th>
<th>Cumulative solar radiation</th>
<th>Wind speed</th>
<th>Greenhouse temperature</th>
<th>Outdoor temperature</th>
<th>Water usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Month</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>-0.002</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Humidity</td>
<td>0.252</td>
<td>-0.169</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solar Radiation</td>
<td>-0.028</td>
<td>0.150</td>
<td>-0.534</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cumulative Solar Radiation</td>
<td>-0.050</td>
<td>0.092</td>
<td>-0.286</td>
<td>-0.055</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wind Speed</td>
<td>-0.111</td>
<td>0.104</td>
<td>-0.061</td>
<td>0.178</td>
<td>-0.246</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Greenhouse Temperature</td>
<td>-0.079</td>
<td>0.138</td>
<td>-0.547</td>
<td>0.617</td>
<td>0.161</td>
<td>0.040</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outdoor Temperature</td>
<td>0.339</td>
<td>0.085</td>
<td>-0.238</td>
<td>0.312</td>
<td>0.370</td>
<td>-0.182</td>
<td>0.521</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Water Usage</td>
<td>0.019</td>
<td>0.062</td>
<td>-0.404</td>
<td>0.753</td>
<td>-0.078</td>
<td>0.097</td>
<td>0.453</td>
<td>0.328</td>
<td>1.000</td>
</tr>
</tbody>
</table>
THE MODEL

The greenhouse water usage behaviour was modeled using the neural network-fitting tool in MATLAB. The network (Figure 4) was trained using Levenberg-Marquardt backpropagation. The network was trained, tested, and validated using 75%, 10%, and 15% of the data, randomly chosen. Randomly chosen data mean that 75% (≈75,457 points) are chosen individually at random with no regard for order, which allows for a broad cross-section of data to be used to train the network without introducing bias created by seasonal trends. The two-layer feed forward network contains eight sigmoid hidden layers and one linear output layer. Neural networks have been found to outperform conventional methods such as regression analysis in water demand forecasting (Jain & Ormsbee 2002; Adamowski & Karapataki 2010). Neural networks are a ‘black box’ method for modeling complex systems. As mentioned previously, the dataset contains many input factors that are collinear. This creates the issue of multicollinearity, which can disrupt the performance and reliability of the model. ANN models deal with multicollinearity in the fact that each input layer of the network is comprised of linear combinations of the inputs of the previous layer, and also in the fact that the output is a function of the sigmoidal functions that involve higher-order interactions of the initial inputs. Because of this overparameterization, the network reduces the impact of multicollinearity but at the expense of interpretability of the underlying weights used in the model (De Veaux & Ungar 1994; Gerth et al. 1994).

GLOBAL SENSITIVITY ANALYSIS

The purpose of performing a SA of a model output is to determine which inputs have the greatest effect on the output. When examining methods for performing SA a distinction is drawn between two different methods of SA, local (LSA) and global (GSA). Saltelli et al. (1999) have provided some insight into their differences. LSA involves varying input factors one at a time while holding other factors fixed and examining the effects on the output. The LSA is undertaken at a central point in the input space which limits the ability to observe effects of interactions between factors since the area of the input space explored is nil. GSA explores all possible input values along the search path and addresses the issue of input interaction by exploring a finite region of the input space by examining the variance of the output averaged across all inputs. There are several methods for executing a GSA, herein the variance-based method of the eFAST is used. The choice of eFAST was based on information found in previous sections and also from results of studies that analyzed convergence of various screening techniques (Vanrolleghem et al. 2015), for which eFAST showed superior performance in terms of computational cost and reliability versus Morris Screening (Morris 1991) and standardized regression coefficients (Saltelli et al. 2008a).

Extended Fourier Amplitude Sensitivity Test

The eFAST proposed by Saltelli et al. (1999) is an extension of the Fourier Amplitude Sensitivity Test (FAST) which was introduced by Cukier et al. (1973). FAST and eFAST are quantitative, variance-based methods for carrying out a GSA, meaning both methods quantify the effect each input factor has on the variance of the output of the model. Equation (1) illustrates the quantification of sensitivity where \( Y \) is the output of the model, \( X \) is an input factor and \( E(Y|X) \) is the expected value of \( Y \) based on a fixed value of \( X \), where \( \text{var}_X \) is taken over all values of \( X \). The advantage of using eFAST over FAST is the latter calculates only first order effects, which do not account for interaction between input factors; eFAST allows for the quantification of first order and total indices, which allows for the calculation of interaction effects.

\[
\text{var}_X[E(Y|X)] \quad \text{var}(Y)
\]

Both GSA methods use sinusoidal functions to create a space filling set of samples for each input factor. To visualize
this process imagine a box containing a sine wave: the box represents the input space containing all possible values of a certain input factor and the sine wave represents the path from which samples or values of the input are being taken. The sine wave can be modified so that it passes through every point in the input space (the box) allowing for a full range of values of each input factor to be sampled. These methods also utilize an expansion of the Fourier series, using Fourier coefficients to estimate the sensitivity of each input factor. In order for FAST and eFAST to be used, a set of transformed input factors must be generated. These transformations require the use of frequencies \( \omega_i \) which must be assigned to each input factor; an algorithm for choosing frequencies is proposed by Saltelli et al. (2008b). The main criterion for choosing frequencies is that they must be incommensurate, meaning they cannot be linear combinations of each other.

\[
x_i = \frac{1}{2} + \frac{1}{\pi} \arcsin (\sin (\omega_i s + \varphi)) \\
-\pi \leq s \leq \pi
\]  

(2)

The transformation used in eFAST is given in Equation (2) where \( x_i \) is the transformed value of the \( i \)th input factor, \( \omega_i \) is the frequency chosen for input \( i \), \( s \) is a set of evenly spaced values chosen between \(-\pi\) and \( \pi \) used for activation, and \( \varphi \) is a random phase-shift used to ensure the sampling curve does not pass through the same points twice and is chosen uniformly between the values 0 and 2\( \pi \). Equation (2) is used for a normalized dataset for which the values of the inputs fall between 1 and 0. Modification of this equation to encompass any input values will be addressed later in this paper. A major issue when dealing with the frequency domain is the Nyquist frequency and aliasing. These issues are dealt with in eFAST by defining a sample size that is sufficiently large.

eFAST first order indices

\[
S_i = \frac{\hat{D}_i}{\hat{D}_{\text{FAST}}} \\

(3)
\]

As described above, the benefit of using eFAST is the ability to quantify the first order and the total sensitivity indices. The first order indices or first order effect is denoted by \( S_i \) (Equation (3)) and is calculated the same way as in FAST, by assigning a unique incommensurate frequency \( \omega_i \) to each input factor, then evaluating the ratio of the variance associated with each frequency \( \hat{D}_i \), Equation (4)) to the total variance of the output \( \hat{D}_{\text{FAST}} \), Equation (5)). This is made possible by using Parseval’s Theorem. The summation of first order indices \( S_i \) for a linear model should be equal to 1, showing that all of the variance of the model output is accounted for without including the effects of interaction. If this is the case, the first order indices are sufficient for calculating the importance, or sensitivity of each factor.

\[
\hat{D}_i = 2 \sum_{p=1}^{M} (A_{s_{p_i}}^2 + B_{s_{p_i}}^2) \\
\hat{D}_{\text{FAST}} = 2 \sum_{p=1}^{\infty} (A_{s_{p}}^2 + B_{s_{p}}^2) \\
\]

(4)

(5)

eFAST total indices

When the summation of the first order effects is not equal to 1 it shows that a portion of the variance of the model output is found in the interaction of the input factors and that the model is non-linear. In order to quantify the importance or sensitivity of each factor in a non-linear model, evaluation of the total indices \( S_{\text{TI}} \) must be performed. Saltelli et al. (1999) proposed a method for evaluating these total order effects by calculating the variance of all factors excluding the input factor being examined \( \hat{D}_{-i} \), where \(-i\) stands for all but the \( i \)th factor and is calculated using Equation (6). This procedure is performed by assigning one frequency to the input factor being examined, and assigning another, much lower frequency to all of the other inputs, and an algorithm for assigning these frequencies is proposed by Saltelli et al. (2008a). This now allows for the calculation of the total effect or total indices of each input factor using Equation (7).

\[
\hat{D}_{-i} = 2 \sum_{p=1}^{M} (A_{s_{p_{-i}}}^2 + B_{s_{p_{-i}}}^2) \\
S_{\text{TI}} = 1 - \frac{\hat{D}_{-i}}{\hat{D}_{\text{FAST}}} \\
\]

(6)

(7)
The procedure for calculating the eFAST sensitivities was carried out using SimLab sensitivity analysis software created by the European Commission Joint Research Council. SimLab implements the procedure outlined in the previous sections and generates an eFAST sample that is unique to each input factor by modifying Equation (4) using the mean and standard deviation of each factor to ensure that the full range of possible values are sampled. SimLab also implements the algorithm for selecting frequencies for each input factor on the basis of a selected sample size, which was chosen to be 1,480 based on the recommendations of Saltelli et al. (1999).

RESULTS AND DISCUSSION

MATLAB neural network model

GSA will produce a set of sensitivities for each input in a given model; this means that in order for the GSA to produce accurate information, the model needs to produce an accurate depiction of the behaviour of the system. Figure 5 displays the error histogram for the MATLAB neural network used in the analysis. Errors are calculated by subtracting the known outputs of the training dataset from those generated by the model. It is noticed that the highest concentration of errors is located near zero, and the mean and root-mean-squared error are \(-0.621 \text{ L} \) and \(642.003 \text{ L} \), respectively. The coefficient of determination \(\left( r^2 \right) \) for the model error is the square of the Pearson product-moment correlation coefficient. For this model \( r^2 \) is equal to 0.712. The average absolute relative error, given by Equation (8), is a statistic used in evaluating the performance of neural networks (Adamowski & Karapataki 2010), where \( O_i \) is the observed or target output, \( F_i \) is the forecast or modeled output and \( N \) is the number of data points. This statistic cannot be used in this case due to the issue of division by zero. The dataset contains a large portion of zero water usage data points, which will cause Equation (8) to approach infinity.

\[
AARE = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{O_i - F_i}{O_i} \right| \times 100\% \tag{8}
\]

Figure 6 contains a plot of the MATLAB neural network output. The dataset starts 1 January 2012 and ends 18 November 2014 and depicts a distinct pattern in water usage. Since there were 100,609 data points encompassing 3 years used in this analysis, it can be estimated that each

---

Figure 5 | MATLAB neural network error histogram.
year contains $100,609/3 \approx 33,500 \ (3.35 \times 10^4)$ data points, which are represented by a solid black line in Figure 6. It can now be seen that each cycle represents 1 year, and that the peak water usage occurs around the mid-point of the cycle, which would correspond to the mid-point of the year, which matches the seasonal pattern shown in Figure 3. Since this pattern is repeated for each of the 3 years it can be said that there is a relationship between time of year and greenhouse water usage. It should be noted that Figure 6 shows negative values across the output space with the largest occurrences appearing at the beginning of 2014. These negative values can be caused by overfitting of the data, which is caused by spurious relationships being drawn in the model that do not exist and create noise in the output signal. Overfitting can be caused by the inclusion of variables that do not in reality have an effect on the output of the model, such as wind speed. The large number of zero water usage data points might also be the cause of the negative consumption values. To rectify this, a floor of 0 could be set in the model to prevent any negative values from occurring. Also, Figure 6 shows peak water usage as $\approx 4,800$ L where as in Figure 3 the maximum is $\approx 7,500$ L. This shows that there are errors in the model in terms of magnitude, but the seasonal patterns have been captured. Overall, this model contained at least one variable (wind speed) that is known to have little to no effect on water usage. This inclusion will likely have a negative effect on model performance and can be used to test the results of the GSA as it should not appear as a highly influential input.

Global sensitivity analysis

The results of the eFAST GSA are shown in Table 3. The summation of the first order indices is 0.7176, which means that approximately 28% of the variance of the model output is not accounted for in the first order indices. Since the first order indices ($S_i$) do not equal 1, the total order indices ($S_T^i$) need to be calculated in order to determine the sensitivities caused by the interaction between inputs. The summation of $S_T^i$ is greater than 1 (1.6169), which also shows that there is cross-correlation between input variables.

Table 3 contains first ($S_i$) and total order ($S_T^i$) sensitivity indices for each factor ranked in order of first order sensitivity. The factors that have the most influence on the variance of the output are time, solar radiation, and outdoor temperature, accounting for 92.4% of the first order indices and 71.7% of the total order indices. As expected, wind ranks low on the first (0.0086) and total order indices (0.0767), but it seems unusual that greenhouse temperature has the lowest first order sensitivity (0.0036), when intuitively it should have a much higher effect. The possible explanation for this low ranking could be overgeneralization.
during the training process as it ranks 8th in first order and 7th in total order where it has similar sensitivity to that of wind. Another possibility is that the greenhouse temperature has no effect on the watering schemes used by the greenhouse, and that the top ranked factors like time and solar radiation are the main drivers. Multicollinearity might be thought to be an issue as greenhouse temperature and solar radiation have a high correlation coefficient (0.617) except that relative humidity also has a high correlation coefficient with solar radiation (~0.534) and it ranked higher in the sensitivity index. This coupled with the findings of Doorenbos & Pruitt (1977), which showed solar radiation as a major factor in plant water need, strengthens the reliability of the results of the GSA. The input factor month ranks 6th in the first order and 5th in the total order indices. This is counterintuitive when Figure 6 shows such a clear yearly trend. This issue might be resolved by using weeks or days in the ANN model in place of decimal months. The reasoning for this is that the narrow range of months (1–12.97) may cause generalization issues when training the ANN as the small changes in the month may be difficult to correlate with water usage, whereas using a larger range may prevent this and generate a higher sensitivity for the seasonal input factor. Another explanation for the low ranking of the month is that all of the seasonal effects are captured within the highly seasonal solar radiation and that the month is not specifically used in the watering schemes. When looking at the total effect it is noticed that the top four factors are in the same order as in the first order index, and the final four factors switch ranking (5 with 6, and 7 with 8). These bottom four factors account for only 20.7% of the total indices and 5.1% of the first order indices.

### CONCLUSION

The results of the GSA provided insights into the factors driving water usage in greenhouses growing bell peppers in southwestern Ontario. For the case studied it was determined that time (decimal hour), solar radiation, and outdoor temperature are the three main factors responsible for the variance in the model output. These factors account for 66.3% of the model variance and 92.4% of the first order sensitivity ($S_i$). The rank of these inputs remains constant through the analysis of the total indices ($ST_i$), for which they account for 71.7%. Inclusion of cumulative solar radiation in the model would increase the total sensitivity accounted for to 79.2%. Including cumulative solar radiation in the model would not require any extra instrumentation since solar radiation is already being monitored. Although time of year, more specifically month, did not rank high in the GSA, it should also be included in the model considering the seasonal variation observed in the output data. The main result of the GSA was justification of the use of only external input factors in a greenhouse water demand model, which is useful to a water utility since they are accessible and are shown to be influential predictors of greenhouse water demand. These external factors can be measured by the water utility by installing metering devices at a centralized location that collects and sends data back to the utility. Since this method is dealing with short-term forecasting (15 minute) the current readings from the telemetry should provide reliable results for forecasting the next period.

This study focused on water usage inside the greenhouse in order to provide a depiction of the watering habits utilized by the growers. Since there is a disparity in technologies implemented in greenhouse operations throughout the world (water recycling, water storage, and climate control) the ability to forecast in-greenhouse water usage will allow for a broader application of the techniques and results of this study. In order to address the issues

### Table 3: First order and total indices for all input factors in order of importance

<table>
<thead>
<tr>
<th>Input factor</th>
<th>First order indices ($S_i$)</th>
<th>Total indices ($ST_i$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>0.4071</td>
<td>0.6385</td>
</tr>
<tr>
<td>Solar Radiation</td>
<td>0.2051</td>
<td>0.3397</td>
</tr>
<tr>
<td>Outdoor Temperature</td>
<td>0.0514</td>
<td>0.1818</td>
</tr>
<tr>
<td>Cumulative Solar Radiation</td>
<td>0.0173</td>
<td>0.1221</td>
</tr>
<tr>
<td>Greenhouse Relative Humidity</td>
<td>0.0141</td>
<td>0.0818</td>
</tr>
<tr>
<td>Month</td>
<td>0.0104</td>
<td>0.0981</td>
</tr>
<tr>
<td>Wind</td>
<td>0.0086</td>
<td>0.0767</td>
</tr>
<tr>
<td>Greenhouse Temperature</td>
<td>0.0036</td>
<td>0.0782</td>
</tr>
<tr>
<td>Σ</td>
<td>0.7176</td>
<td>1.6169</td>
</tr>
</tbody>
</table>
related to technology and alternative water supply, this process can be refined by installing metering on the water lines feeding the greenhouse operation in order to capture the amount of water drawn from the water utility. This utility or fresh water demand can then be subtracted from the greenhouse demand modeled using the results of this study, leaving only the recycled water usage and providing a full picture of greenhouse water use.

This method can be applied to various crop types such as tomatoes, cucumbers, melons and flowers, with the goal of developing a comprehensive view of greenhouse water demand. These results can be implemented in a piecewise water demand-forecasting model that will address the unique combination of greenhouse crops at any given location. Using this technique, greenhouse operators can expose underlying patterns in the manual watering schemes utilized by many growers in order to work towards full automation of crop watering.

ACKNOWLEDGEMENTS

The authors would like to thank Union Water Supply System, Ontario Clean Water Agency, and Ontario Greenhouse Vegetable Growers for their continued support of this research.

REFERENCES


Ma, L., He, C. & Wang, Z. 2013 The research for the greenhouse water evaporation based on the environmental factors. Advance Journal of Food Science and Technology 5 (8), 1049–1054.


First received 26 October 2015; accepted in revised form 22 February 2016. Available online 19 March 2016