

**Municipal and industrial water withdrawal
modelling for the years 2000 and 2005
using statistical methods**

February 2011



This document describes the efforts to generate models that estimate the municipal and industrial water withdrawals for the years 2000 and 2005.

Table of contents

1	Introduction.....	3
2	The WaterGAP2 model and its applicability to AQUASTAT	3
2.1	Municipal water withdrawal.....	5
2.2	Industrial water withdrawal	6
3	AQUASTAT modelling: methodology and analysis	7
3.1	Data analysis.....	7
3.2	Model.....	7
3.2.1	Modelling approach: strengths and weaknesses	9
3.2.2	Dummy variable analysis	10
3.2.3	Independent variables considered	10
3.2.4	Transformations	11
4	Results and conclusion.....	11
	Appendix A - WaterGAP 2 statement	14
	Appendix B - WaterGAP 2 industrial water withdrawal analysis	15

List of figures

Figure 1 - Conceptual model of structural change in the domestic (a) and industrial (b) sectors.	4
Figure 2 Plot of country-level data vs. MWSI sigmoid curve for all countries using the best-fit gamma coefficient = $3.28E-14$	5
Figure 3 Plot of country-level data vs. IWSI curve for all countries using the best-fit gamma coefficient = $1.00E+7$	6

1 Introduction

AQUASTAT is FAO's global information system on water and agriculture developed by the Land and Water Division of the Food and Agriculture Organization (FAO) of the United Nations. It provides a variety of topical studies on water use and irrigation, but the online database is AQUASTAT's flagship product. The amount of information provided on the state of water resources and uses is frequently updated, and unparalleled. Due to this, the World Water Development Reports, UN Water, the World's Water reports, as well as several international and national water experts, researchers and policy-makers rely on the data in AQUASTAT's main country database for their analyses and decision-making.

AQUASTAT is, and will remain, a database that documents and reports statistics reported by countries. However, data gaps are unavoidable due to lack of resources and lack of capacity (or willingness to share data) at the country level. In order to increase the usability of the AQUASTAT database, some limited and careful modelling is necessary. This exercise was carried out to supplement the AQUASTAT database with municipal and industrial water withdrawal data values for 2000 and 2005. The modelled data in the database can be identified by the modelled data qualifier. Of course, modelled data should not be included in subsequent modelling or numerical methods, as the uncertainty would be compounded further with additional modelling. It is up to the end user to use modelled data responsibly. Additionally, the constraints and applicability notes of this document should be well taken into consideration, since modelled data is always susceptible to some error.

Water withdrawal patterns have been observed and modelled by many. It is apparent that water extraction patterns correlate to a country's affluence, technology, population and relative importance of the different sectors. Several researchers have opted to use simple proxies or correction factors in mathematical calculations to estimate water withdrawal (Döll and Hauschild, 2000; Fujino, 2001). However, the relationship between water withdrawal and socioeconomic characteristics is quite complicated. In order to be able to assess the validity of simple proxy relationships, it is important to understand the theory behind water withdrawals and how they change as a function of other variables. Several researchers have tackled this problem theoretically (Shiklomanov, 2000; Alcamo *et al*, 2003), and using empirical relationships. Of these, the model that we found closest matched the AQUASTAT data was the WaterGAP2 model. This model is presented in the next section in order to evaluate its applicability for AQUASTAT to fill data gaps in its municipal and industrial water withdrawal data. Section 3 describes the final methodology used by AQUASTAT to estimate municipal and industrial water withdrawal data.

2 The WaterGAP2 model and its applicability to AQUASTAT

A brief explanation of municipal water withdrawal changes is provided in the WaterGAP 2 model documentation, presented hereunder (the use of the term

domestic water withdrawal is closely related to AQUASTAT's municipal water withdrawal):

[...] “structural change” is the change in water intensity that follows from a change in the structure of water use (that is, the combination of water-using activities and habits within a sector). In the domestic sector, poorer households with sparse indoor plumbing may gradually acquire more water-using appliances as their income increases. Eventually, the average household becomes saturated with dishwashers, washing machines, and other water-using appliances, and as a result water use stabilizes. The consequence of these structural changes is that average water intensity of households (m^3 person $^{-1}$) first sharply grows along with the growth in national income, but eventually stabilizes as national income continues to grow”. (Alcamo et al, 2003)

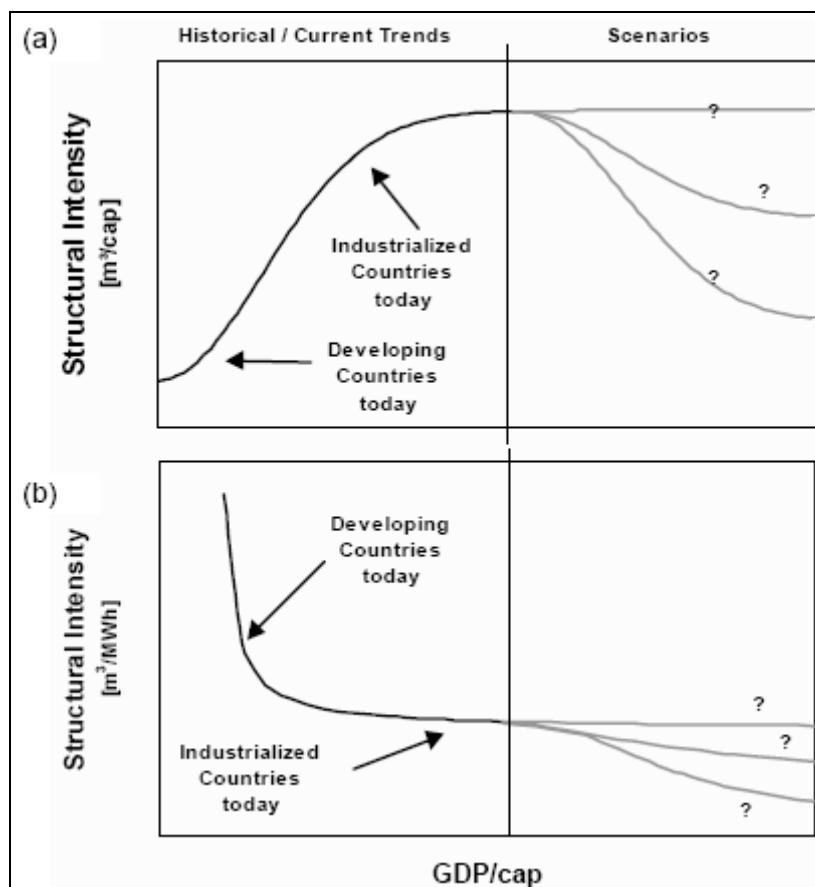


Figure 1 - Conceptual model of structural change in the domestic (a) and industrial (b) sectors (Source: Alcamo et al, 2003)

This model was developed for sub-national datasets and it is therefore necessary to verify the model's applicability to national-level data as used in AQUASTAT. Admittedly, national-level withdrawal information does dilute the usability of technical information as the country average value is not necessarily representative of any specific region in a country. While AQUASTAT does count with some sub-national level data, most countries report only country-wide averages.

2.1 Municipal water withdrawal

Figure 2 shows the relationship between the country level municipal water withdrawal AQUASTAT data vs. the prediction of the WaterGAP 2 Domestic Water withdrawal model mentioned above (NOTE: The word “prediction” and words derived from it and used in this article do not mean future-looking, but refer to the results of the models). This figure was prepared by using the equations provided by WaterGAP 2, included in Appendix A. Clearly, there is no correlation, when considering the global picture, which renders this model of little use for this particular analysis. While it is true that the model is designed for domestic water withdrawals, whereas AQUASTAT reports municipal water withdrawals, this is not expected to change the result drastically as the two are correlated. The fact that there is little concordance between the observed model and the country-level data from AQUASTAT is not a criticism of either the model or the data, but simply means that this approach is not useful when analyzing aggregate country-level information at a global scale.

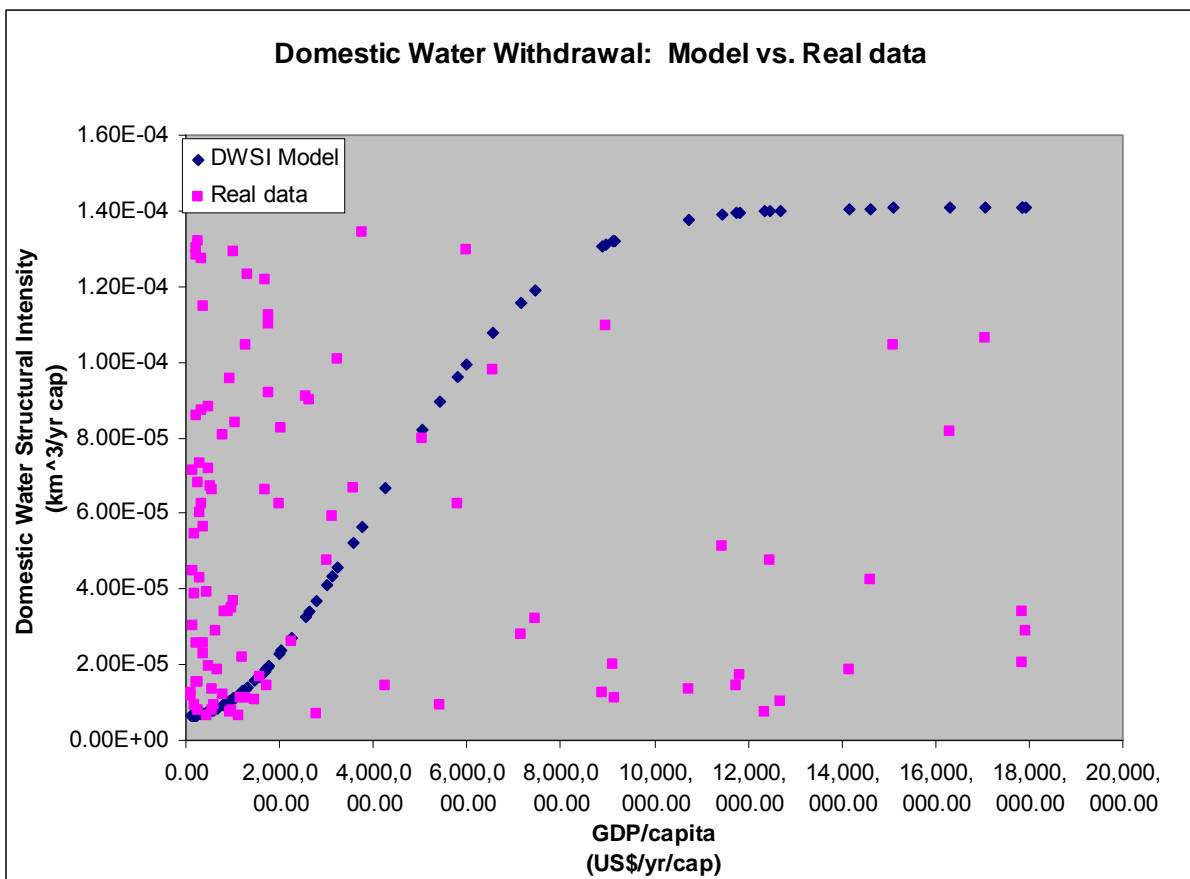


Figure 2 Plot of country-level data vs. MWSI sigmoid curve for all countries using the best-fit gamma coefficient = $3.28\text{E-}14$

While this model did not closely match the global picture as per AQUASTAT data, its utility for analyzing the behaviour of individual countries was also evaluated. Unfortunately, in AQUASTAT only a few countries have sufficient data to allow for a long-term trend analysis, but this analysis was performed where possible, and there doesn't appear to be conclusive evidence that individual country data follows the relationship stated in the WaterGAP 2 model.

2.2 Industrial water withdrawal

The industrial model does seem to fit the global dataset better, as can be observed in Figure 3. Statistically speaking, however, the vast majority of the data is not explained by the predicted relationship (notice the strong grouping of data-points closer to the origin). Appendix B provides the mathematical equations used to generate this graph, and the resulting statistical analysis. Individual countries were also compared against this relationship, with mixed but inconclusive results.

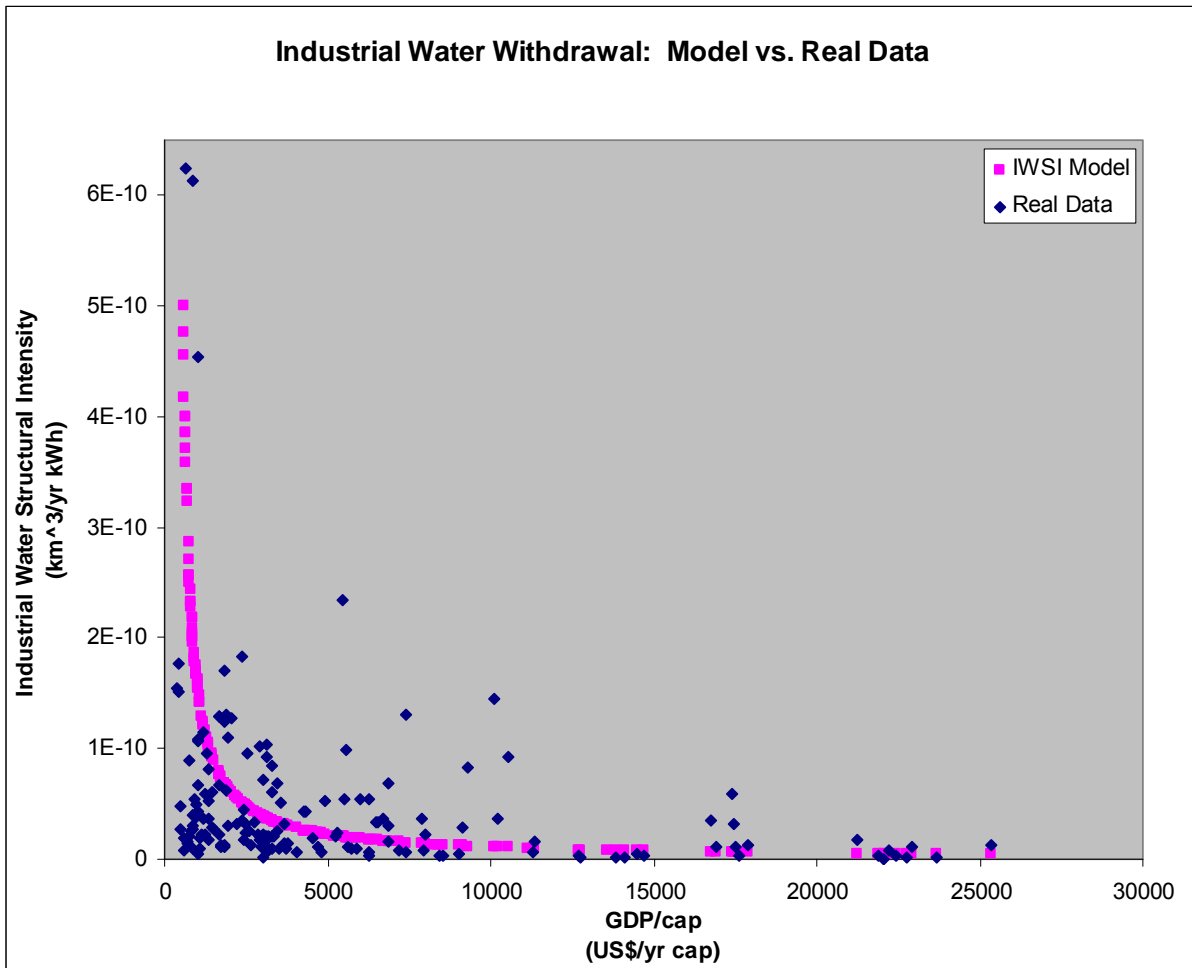


Figure 3 Plot of country-level data vs. IWSI curve for all countries using the best-fit gamma coefficient = 1.00E+7

In the absence of a theory-based approach that is valid and relevant to the AQUASTAT dataset, statistical regression analysis was used to model withdrawal patterns, based on information that was known. In this way, empirical relationships were identified, even though these solutions are highly specific to this exercise, and not representative of extrapolatable relationships.

3 AQUASTAT modelling: methodology and analysis

3.1 Data analysis

In AQUASTAT, industrial water withdrawal refers to self-abstracted water, while municipal water withdrawal represents water provided by public networks. These networks mainly provide water to households (domestic water use), but can also provide water to other urban services, including stores, markets, tourism centres, and urban industry. Frequently, however, data is not provided with a level of detail that would allow for subsequent disaggregation. Some unknown percent of municipal water is used in industrial activities which, depending on the country, may be reported under municipal or industrial water withdrawals. In order to deal with this data problem, the municipal and industrial water withdrawals were added and modelled together, and subsequently disaggregated again based on the observed ratios in that country, and its neighbours.

AQUASTAT data exists in three dimensions: country, variable, and year. Statistical analyses that cover three dimensions like Panel analysis, are quite intricate, and therefore require a significant time investment to accomplish. In order to facilitate the analysis, countries and years were combined into unique country-year entries. Box 1 illustrates the two-dimensional representation of the three dimensional data that exists in AQUASTAT. This is a valid simplification as long as tests are conducted to evaluate if there are period-wide effects that would affect withdrawal values during a certain period (see Section 3.2.2 for controls). One potential issue with this modelling exercise is that AQUASTAT has several withdrawal values for some countries, and very few for others. This does create a certain bias in the modelling exercise, but this is unavoidable without losing a significant amount of data.

Box 1 – Two-dimensional solution to three-dimensional dataset.

<u>Values of industrial water withdrawal:</u>	<u>Change to:</u>
1992 Country X = 2.3 Country Y = 3.1	Country X - 1992 = 2.3 Country Y - 1992 = 3.1
1993 Country X = 2.4 Country Y = 3.1	Country X - 1993 = 2.4 Country Y - 1993 = 3.1
1994 Country X = 2.6 Country Y = 3.2	Country X - 1994 = 2.6 Country Y - 1994 = 3.2

3.2 Model

The inputs to this model are:

- Independent variables: For each model, the predictive value of several potentially predictive variables was tested using Ordinary Linear Regression.

- Dependent variables: Municipal and industrial water withdrawals in the AQUASTAT database from 1960-2010 and from all countries were considered for this analysis. Exceptions: OECD country data and previously modelled data was not considered.

The goal of this document is to strike a balance between simplicity and thoroughness. It is paramount that the users of the AQUASTAT database understand the model in order to understand the limitations of this model.

The modelling option selected was simple (ordinary) linear regression, where a dependent variable (in this case, our water withdrawal variables) can be predicted using a variety of independent variables. This approach is also referred to as Multivariate analysis. The model takes the form:

$$Y_t = \beta_0 + \beta_1 * X_{1t} + \beta_2 * X_{2t} + \beta_2 * X_{2t} + [...] + e \quad [\text{EQ 1}]$$

Where Y is the dependant variable, each X is an independent variable, β_0 is the intercept, $\beta_1, \beta_2, [...]$ are the regression coefficients for each independent variable, t represents each case analyzed, and e corresponds to the error of this linearization. We have to assume that the error is independent of X, normally distributed, and has constant variance (Cook and Weisberg, 1999). Explained in simple terms, we have to check to make sure that the model doesn't fit very well in one part of the dataset and very poorly elsewhere.

Care was taken to select variables that were expected to be linked to the water withdrawals. For example, it makes sense to expect that municipal water withdrawal somehow is tied to the number of people that use water (population). Of course, there are several different variables tied to the number of people that exist in a given country, for example: total population, urban population, rural population, total labour force, number of people within different age, etcetera. In the effort to determine the best model that works for a global dataset and across all different periods, several of these variables were tested to find the best possible relationship.

The statistical package Arc (XLISP-STAT) was used to perform the analysis. The statistics considered in the analysis include the r^2 value, t-test and F tests, and p-value. The r^2 and F values provide insight on the quality of each model. The r^2 indicates the overall predictive power of the model. The t-test determines the predictive power of each variable, and is used to decide which variables are most significant. The p-value, also known as the α -value, indicates the probability that the variable in question is NOT significant. p-value of less than 0.05 means that there is a 5 percent chance or less that the variable is not significant in determining the model outcome. Given the number of data-points used, a t-test value of 2.0 was generally comparable to a p-value of 0.05 (which means there is only a 5 percent chance that the indicator is not predictive).

An iterative approach was used to determine the best model, using the r^2 and F values to indicate if the iterations were more or less predictive. Also, weighting by Area or Gross Domestic Product per person (GDP/cap) improved the regression predictions, and all models were weighted by one of these two variables.

Naturally, water withdrawal profiles for all the different countries around the world vary considerably. In generating regional groupings, it was observed that OECD countries frequently had municipal and industrial water withdrawals that were inconsistent with the other countries in their respective regions. Due to this fact, OECD countries were not included in this modelling exercise (data for OECD countries already exists, therefore modelling of these countries was not required). The need to break up the global dataset into more manageable groups led to the formulation of the following groups, which were each analysed individually:

- Europe
- Middle East
- India & China
- Asia *
- Latin America and the Caribbean
- Africa

* Asia does not include Middle East countries, India, or China. These countries were found to have withdrawal pattern different enough to justify their own models.

Other groupings were also analyzed but were found to provide inferior fit results.

3.2.1 Modelling approach: strengths and weaknesses

This approach presents certain strengths, but is not without weaknesses. For transparency and simplicity, these are listed below in bullet format.

Strengths

- This approach estimates withdrawals based on time series, not based on one year.
- The “business-as-usual” scenario is not assumed, instead, the data itself determines the relative increase or decrease of values for the target years.
- By modelling municipal and industrial water withdrawals together, some tolerance to reporting inconsistencies in nationally reported data is afforded.

Weaknesses

- Analysis was conducted at a continental basis (Africa, Asia, Latin America and the Caribbean, Europe), making exceptions for OECD countries, India, China, and the Middle East regions. As expected, water withdrawal patterns within any given country do not necessarily reflect the regional/continental withdrawal patterns.
- Countries that have reported data for several years affect the model more than countries with few data-points,. Still, this bias was seen as less bad than discarding excess data from well-reported countries, as this could introduce a bias depending on the year reported, and would drastically diminish the number of observations on which the model could be based.
- The data taken into consideration is that which has been reported by individual countries. While this data has passed rigorous quality assurance by AQUASTAT, there still is the possibility that any given number is not representative of the real situation.

- Only independent variables with full data sets could be included in this analysis. Variables which were seen as highly explicative could not be used as they would not help calculate withdrawals for the target years due to significant data gaps.

3.2.2 Dummy variable analysis

The condensing step of flattening the dataset into country-year entries assumed that there were no intrinsic differences based on time or space. For example, perhaps the withdrawal numbers for Asian countries from 1975 were in general lower than the rest of the dataset. In order to test this assumption, a dummy variable analysis was done to determine if there were particularly influential periods or geographical groupings.

Analysis of different periods and regional effects showed that Asia and 2010 had results significantly different from the rest of the dataset. Both can be explained by the fact that only a few data-points are available for 2010, mostly in Asia, and are relatively large in magnitude. Understanding this, the error was not expected to be significant at a global level. That being said, the information about the potential error due to the aggregation methodology employed was introduced in the model, instead of rejecting the flattening approach outright. This was accomplished by introducing the relevant dummy variables into the model. In this way, the potential weakness of the model could be explained and therefore accounted for. As expected, the dummy variables did not provide enough additional explicative power to justify their presence in the model, and were therefore removed.

3.2.3 Independent variables considered

The types of variables considered include water resources, area, population, electricity consumption (as a proxy for industrialization), state of economy, and labour force. It is immediately apparent that several of these variables, while interesting, simply do not have enough information to be useful. The labour force was particularly predictive, however due to data gaps in that dataset, none of the labour indicators were included in the final model. As mentioned earlier, an iterative approach was used to determine which of these predictors was most significant. Variables with p -values of less than 0.05 were considered to be appropriate.

The final variables selected were total population (UNPD, 2008), and total electricity consumption (US-EIA, 2008). While some amount of co-linearity between these two variables can be intuitively expected, in aggregating all the different countries no significant co-linearity was observed in the transformed variables during the analysis.

The electricity consumption values were generated by summing “Conventional”, “Nuclear” and “Other” energy classes. In light of the fact that a substantial portion of the water used for cooling of energy plants is saline and thus outside the purview of this report, the energy consumption numbers were reduced according to each countries shoreline. GIS was used to generate an estimate correlating the amount of cooling water that occurs near coastlines with known saline water withdrawals for

energy generation. In this way it was assumed that water withdrawals that occur up to 50 km from each coast use saline water for energy generation. This was only found to make a substantial difference in the Latin American region (presumably due to the large number of countries with a substantial amount of their withdrawals within 50 km of the coastline).

3.2.4 Transformations

Transformations change the data in a way that maximizes its usability. For example, instead of using GDP, a better data descriptor might be $GDP^{-1.5}$ or $\log(GDP)$. This maximizes the descriptive power of each independent variable within the model being analysed. Other benefits include:

- Normalizing the data (reducing the differences in the dataset). Depending on its size, population, etc., the withdrawal numbers vary considerably from country to country, which makes plotting of the data difficult. In these cases, the data can be ‘compressed’ by using the log function. Of course, this benefit is only for presentation purposes, not used in the model per se.
- As mentioned previously, the models are strongest when the variance is constant and normally distributed. The ‘Box-Cox’ is an automated method for transforming data that analyzes the variables in use and finds the transformations that maximizes the normality of the variance (Cook and Weisberg, 1999). This method of transforming variables is preferable to attempting to find optimal transformations manually, as it is quicker and provides better results. Therefore, this transformation methodology was used in the models..

4 Results and conclusion

AQUASTAT used statistical software to model municipal and industrial water withdrawals together, and then disaggregated the predicted value according to known ratios of municipal/industrial withdrawals. Since this model is empirical in nature, this model is not forward-looking, i.e. – it is not designed to predict future values, and should not be used as such.

The following statistical results were obtained from this modelling exercise, provided here as an indication of how well the estimated values fit the rest of the dataset:

Table 1 – Regression results

	r^2	F	Weight
Middle East	0.907	286.29	GDP/cap
Europe **	0.978	916.35	GDP/cap
India & China	0.998	1718.62	GDP/cap
Asia	0.846	66.35	Area
Latin America and the Caribbean	0.978	476.50	Area
Africa	0.920	427.33	GDP/cap

** During the analysis, it was seen that a better fit could be accomplished by dropping energy Consumption from the analysis, as its t-value and corresponding p-value were above the 5 percent tolerance level. Therefore this model depends only on population, and not on energy consumption.

After obtaining acceptable modelling results, each model was thoroughly tested in order to ensure the model was acceptable for use. The t-test directionality (positive or negative) was checked to ensure that the relationship determined by the model did make sense in reality, and did not merely represent a mathematical coincidence.

Residual is the difference between a data-point and what the model predicts that value should be. Residual plots were constructed and compared along the range of the model. This was done to identify if any results looked particularly poor, and also to ensure that the residual was normally distributed.

The Cook's Distance is an indicator showing how much influence any given data-point has over the entire model. If the Cook's Distance is large for a point that also has a large residual, then this means that the model is very much "trying" to explain that data-point, but is failing to do so correctly. This might indicate that the model is not very good for this particular data-point. The points that had the highest Cook's Distance were analyzed during the quality control step, and also led to splitting Asia into sub-regions..

After the model results were disaggregated back into municipal and industrial water withdrawals, each country's results were evaluated individually. In this way, countries with either substantially ill-fitting results or countries whose withdrawal patterns differed from their continental grouping could be eliminated before inclusion in the database.

As a final comment, the AQUASTAT staff would like to emphasize the fact that this modelling exercise is neither the norm of the programme or intended to be used for any purpose other than what has been stated. This exercise has been carried out to provide a point of comparison in order to supplement the information already in the database. These modelled results are likely to be updated as soon as more reliable data is available.

References

Alcamo, Joseph; Döll, Petra; Henrichs, Thomas; Kaspar, Frank; Lehner, Bernhard; Rösch, Thomas; Siebert, Stefan. 2003. Development and testing of the WaterGAP 2 global model of water use and availability. *Hydrological Sciences Journal*, 48(3).

Cook, R. D. and Weisberg, S. 1999. *Applied Regression Including Computing and Graphics*. Wiley, New York.

Döll, P. and Hauschild, M. 2000. *Large Scale Water Use Model NoWUM for modelling water management in Ceará and Piauí in northeast of Brazil*. University of Kassel, Kassel, Germany.

Fujino, J. 2001. *AIM/TREND MODEL*. National Institute for Environmental Studies, Tsukuba, Japan

Shiklomanov, I. 2000. *Appraisal and assessment of world water resources*. *Water Int.*25, 11–32.

UNPD [United Nations Population Division]. 2008. *Long term series estimates and projects from 1961 to 2050*. UN Revision 2008. Through FAOSTAT <http://faostat.fao.org/site/452/default.aspx>, accessed on 15/12/2010.

US-EIA [United States - Energy Information Administration]. 2008. *International Energy Annual 2006*. <http://faostat.fao.org/site/452/default.aspx>, accessed on 15/12/2010.

This report was prepared by Amit Kohli with inputs from Karen Frenken, Ricardo Sibirian, and Antonio Vezzani, whose guidance was pivotal in the generation of this exercise.

Appendix A - WaterGAP 2 statement

From WaterGAP 2 model

Municipal model:

$$MSWI = MSWI_{\min} + MSWI_{\max} (1 - \exp(-\gamma_d * (GDP / cap)^2))$$

Where:

$MSWI =$	Municipal Structural Water Intensity evaluated as municipal water withdrawal over total population
$MSWI_{\min} =$	Minimum value of MSWI
$MSWI_{\max} =$	Maximum value of MSWI
$\gamma_d =$	Curve parameter, determined iteratively to optimally fit this dataset.
$GDP/cap =$	Gross Domestic Product per capita

Industrial model:

$$ISWI = \frac{1}{\gamma_i (GDP_cap - GDP_cap_{\min})} + ISWI_{\min}$$

Where:

$ISWI =$	Industrial Structural Water Intensity evaluated as industrial water withdrawal over total electricity production
$\gamma_i =$	Curve parameter, determined iteratively to optimally fit this dataset.
$GDP/cap =$	Gross Domestic Product per capita
$GDP/cap_{\min} =$	Minimum value of GDP/cap
$ISWI_{\min} =$	Minimum value of ISWI

Appendix B - WaterGAP 2 industrial water withdrawal analysis

Starting from the equation:

$$ISWI = \frac{1}{\gamma_i(GDP_cap - GDP_cap_{min})} + ISWI_{min} \quad [1]$$

This relationship was simplified by normalizing the values ISWI and GDP_cap (by making implicit to the calculation the minimum values), as follows:

$$(ISWI - ISWI_{min}) = \frac{1}{\gamma_i(GDP_cap - GDP_cap_{min})} \quad [2]$$

where:

$$ISWI' = (ISWI - ISWI_{min}) \quad [3]$$

$$GDP_cap' = (GDP_cap - GDP_cap_{min}) \quad [4]$$

therefore:

$$ISWI' = \frac{1}{\gamma_i(GDP_cap')} \quad [5]$$

$$ISWI' = \frac{1}{\gamma_i} * \frac{1}{GDP_cap'} \quad [6]$$

$$ISWI' = \delta * GDP_cap'^{-1} \quad [7]$$

We now have a linear relationship from which δ can be empirically deduced. The following analysis confirms numerically that this model is not highly descriptive of the dataset.

Data set = Industrial_WW-Int_mins, Name of Fit = L1

Normal Regression

Kernel mean function = Identity

Response = y'

Terms = (x'^{-1})

With no intercept.

Coefficient Estimates

Label	Estimate	Std. Error	t-value	p-value
x'^{-1}	1049.50	226.632	4.631	0.0000

Sigma hat: 9.17296

Number of cases: 160

Degrees of freedom: 159

Summary Analysis of Variance Table

Source	df	SS	MS	F	p-value
Regression	1	1804.44	1804.44	21.44	0.0000
Residual	159	13378.8	84.1432		
Lack of fit	144	11332.1	78.6948	0.58	0.9496
Pure Error	15	2046.71	136.447		