# MODELING UNCERTAINTY:
# QUICKSAND FOR WATER TEMPERATURE MODELING
**10/28/2002**

John M. Bartholow
US Geological Survey
Fort Collins Science Center
2150 Centre Avenue, Bldg C
Fort Collins, CO 80526-8118
970-226-9319
John_Bartholow@USGS.Gov

## Abstract

Uncertainty has been a hot topic relative to science generally, and modeling specifically. Modeling uncertainty comes in various forms: measured data, limited model domain, model parameter estimation, model structure, sensitivity to inputs, modelers themselves, and users of the results. This paper will address important components of uncertainty in modeling water temperatures, and discuss several areas that need attention as the modeling community grapples with how to incorporate uncertainty into modeling without getting stuck in the quicksand that prevents constructive contributions to policy making. The material, and in particular the references, are meant to supplement the presentation given at this conference.

Key Words: water temperature, modeling, uncertainty, SSTEMP

*This paper is a slightly revised version of a presentation given at the American Institute of Hydrology Conference on Hydrologic Extremes: Challenges for Science and Management, Portland, OR, October 13-17, 2002.*

## The Problem

*Nondeterminism means never having to say you are wrong.* - Anonymous

Water quality modeling has been and will continue to be important in a variety of assessments of stream health. Recently, the Total Maximum Daily Load (TMDL) requirements of the Clean Water Act have served to increase the level of scrutiny given to water pollution modeling. Water temperature is a key element for many water quality studies and sometimes the only attribute considered for analysis for some streams. The question arises as to how good – how accurate – our models are. This paper introduces many concepts that are important to consider when dealing explicitly with uncertainty in water temperature modeling. It draws from an extensive literature base and serves to pass along references to important works that readers may wish to explore as they familiarize themselves with this topic. This paper also introduces new features that have recently been added to the Stream Segment Temperature Model (SSTEMP; Bartholow

2002) that deal explicitly with how to incorporate measurement uncertainty in modeling. Important unanswered questions are highlighted.

The paper is loosely organized in five sections paralleling the translation of the real world to the policy world as portrayed in Figure 1.

## The Real World

*As far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality.* – Albert Einstein

The real world is fraught with uncertainty through time and space (Cleaves 1994; Essig 1998), and the whole process of translating the real world into a mathematical caricature from which we may learn something is a challenging task. Yet we also know that well-structured models offer numerous advantages in organizing our thoughts and guiding our decisions. Increased understanding of physical systems has produced some excellent models that perform well in terms of reproducing reality. Modeling water temperature specifically is a good example of the progress made in both model development (e.g., Theurer 1984; Brown and Barnwell 1987; Sullivan et al. 1990; ODEQ 1999) and model applications (e.g., BioAnalysists Inc. 2002; Brown ND; Borsuk, Stow, and Reckhow 2001; Rounds, Wood, and Lynch 1998).

## Observations

*Everything is vague to a degree you do not realize till you have tried to make it precise.* - Bertrand Russell

To model successfully, we must carefully observe the real world. We measure, we quantify, we estimate. But measurements only approach the truth; we must always contend with bias and blunder, extrapolation and error. Many authors have commented extensively on the sources of measurement and modeling error/uncertainty, and to some degree how to deal with them (e.g., Suter, Barnthouse, and O'Neill 1987). To summarize, *measurement error* is composed of (1) equipment failures and their inherent accuracy, (2) human error, (3) and natural variability through space and time (Cleaves 1984), i.e., we measure in one place and assume that our measurements are widely representative. We also make errors in artificially trimming the boundary of the system we are looking at, again either in time (because of limited data) or space (because we are ignorant of the full nature of cross-boundary dependencies). *Modeling error*, in contrast, is composed of errors in estimating internal model parameters as well as the inescapable simplifications and assumptions: we extrapolate parameter values from other locations or times, or we borrow them from other models; and models suffer from structural weaknesses either because we have aggregated sub-components that should not have been combined, or inserted elements that we should have omitted because they add unwarranted complexity (Suter, Barnthouse, and O'Neill 1987; Borsuk, Stow, and Reckhow 2001; Brown and Barnwell 1987; Cleaves 1994). We seldom truly know how to improve on what we have done. Finally, there are always elements of surprise that simply push us beyond the boundaries of what we, and our models, 'expect'.

## Modeling Realm

*Far better an approximate answer to the right question, than the exact answer to the wrong question, which can always be made precise.* - John Tukey

We are generally familiar with the trials and tribulations of modeling – calibration and validation – and I will not go into those details here. But there remains a troubling area that I do not think we have adequately thought out. Consider a classic display of goodness-of-fit such as that shown in Figure 2. This chart displays a multi-year fit for mean daily water temperature on the Klamath River at a specific location using the US Army Corps of Engineers physically-based HEC-5Q model. The multi-year sinusoidal form shown in the figure could just as easily be an hourly model. The coefficient of determination between the simulated and observed values is quite high ($r^2 = 0.98$), signaling an excellent goodness-of-fit. But if we ask a slightly different question, "How much does the model explain deviations from the historical normal for each day (or each hour)", we get a far different result as shown in Figure 3. In other words, we are now asking a new question – "How well does the model describe day-to-day (or hour-to-hour) fluctuations from the average?" The same calibrated model, used for a different purpose, now has a coefficient of determination ($r^2$) of 0.52 because, in effect, we have removed autocorrelation from the statistical calculations. This is troubling to me and casts doubt on the confidence in the modeling we have traditionally done for answering certain specific questions.

## Simulated World

*When one admits that nothing is certain, one must, I think, also add that some things are more nearly certain than others.* - Bertrand Russell

Previous versions of SSTEMP were deterministic; you supplied the 'most likely' estimate of input variables and the model predicted the 'most likely' thermal response. This approach was comforting and easy to understand, but we know there is variability in the natural system and inherent inaccuracy in the model. The previous model did not reflect variance in measured or estimated input variables (e.g., air temperature, streamflow, stream width) or parameter values (e.g., Bowen ratio, specific gravity of water); therefore it could not be used to estimate the uncertainty in predicted temperatures. I recently added a feature that may be useful in estimating measurement uncertainty inherent in predicting water temperatures (Bartholow 2002).

The built-in uncertainty routine uses Monte Carlo analysis, a technique that gets its name from the seventeenth century study of the games of chance. The basic idea behind Monte Carlo analysis is that model input values are randomly selected from a distribution that describes the set of possible values composing the input. That is, instead of choosing one value for mean daily air temperature, the model is repeatedly run with several randomly selected estimates for air temperature in combination with random selections for all other relevant input values. The distribution of input values may be thought of as representing the variability in measurement and extrapolation error, estimation error, and a degree of spatial and temporal variability throughout the landscape. In other words, we may measure a single value for an input variable, but we know that our instruments are inaccurate and we also know that the values we measure might have been different if we had measured in a different location along or across the stream, or on a different day.

SSTEMP is fairly crude in its method of creating a distribution for each input variable (see Figure 4). There are two approaches in this software: a percentage deviation and an absolute deviation. The *percentage deviation* is useful for variables commonly considered to be reliable only within a percentage difference. For example, USGS commonly describes stream flow as being accurate plus or minus 10%. The *absolute deviation*, as the name implies, allows entry of deviation values in the same units as the variable (and always in international units). A common example would be water temperature where we estimate our ability to measure temperature plus or minus approximately 0.2 degrees. In both cases, a sample value is chosen from either 1) a uniform (rectangular) distribution plus or minus the percent deviation, or 2) a normal (bell-shaped) distribution with its mean equal to the original value and its standard deviation equal to 1.96 times the deviation entered so that the set of samples drawn represent about 95% of the full distribution. No attempt has been made to account for correlation among variables, even though we know there is some.

SSTEMP's random sampling is used to estimate the average temperature response, both for mean daily and maximum daily temperature, and to estimate the entire dispersion in predicted temperatures. You tell the program how many *trials* to run and how many random *samples* per trial. The standard Monte Carlo approach simply runs many random samples, but by adding another layer -- the trials -- SSTEMP gains a two-fold advantage (Macaluso 1983, 1984). First, by computing the average of the trial means, it allows a better, tighter estimate of the mean simulated value. This is analogous to performing numerous modeling 'experiments' each with the same number of data points used for calibration. Each 'experiment' produces a single estimate of the mean, and the cumulative set of experiments produces a refined estimate of the mean. Second, one can gain insight as to the narrowness of the confidence interval around the mean depending on how many samples there are per trial. The lesson to be learned here is that the number of data points you have to calibrate the model has a profound effect on the narrowness of the confidence interval. For example, if you have only a few days' worth of measurements, your confidence interval will be far broader than if you had several months' worth of daily values.

Once the analysis is complete, a summary of the temperature output is given (also Figure 4). The estimates of the mean are accompanied by the best estimate of their standard deviation and 95% confidence interval, followed by the 'full' estimate of the standard deviation for the complete range of model predictions. Unlike the estimate of the mean and its deviation, the estimate of the 'full' model uncertainty is considerably more broad. The program also supplies an exceedence table displaying the probabilities of equaling or exceeding the stated temperature. Finally, you may plot a bar graph showing the frequency of trial-average results (Figure 5).

As you have seen, SSTEMP is attempting to answer two distinctly different questions: (1) What is the best estimate of the mean, and (2) what is the best estimate of the full model uncertainty. Which one should you use? If your goal is to estimate the mean temperature, I recommend the 95% confidence interval around the mean. This would be 1.96 times the SD of the estimate of the mean, 0.34°F in the example shown in Figure 4. If you want to estimate the variability in the full model predictions, use 1.96 times the full distribution value, 1.21°F in the example. As you can see, these two estimates can be widely different, though this depends on the number of trials and samples per trial. Remember that there is no magic in these statistics; they simply characterize the distributions of the model results.

## Policy World

*All models are wrong, but some are useful.* – George Box

Before we apply uncertainty in the world of policy, it is important to put it into perspective. There are many reasons why models are not used or used poorly (Bunnell 1989). Models may be misused due to misunderstandings of the situation or results. Some applications may be hastily done without proper validation. Models today are increasingly challenged legally, or perhaps branded as "junk science" (Cleaves 1984; Steel et al. 2001). Models can be disused because decision makers really didn't want to use a model at all, it didn't meet the problem objectives, it is too complex or data intensive, or finally, because it is too inaccurate. So you see that inaccuracy is really only a small part of this overall problem and 'correcting' our analyses to add uncertainty is no guarantee of success. We must also keep in mind that incorporating uncertainty can be problematic in the sense that it can be ignored in favor of 'confident experts', swept under the rug by those who don't want to see it, can stimulate procrastination until certainty is achieved (which it can not be), or result in 'overprotective' safeguards that cannot withstand benefit:cost scrutiny (Cleaves 1994). Some have also complained that presenting uncertainty information only serves to provide 'ammunition' for the nay-sayers.

On the other hand, developing one's perspective includes appreciating the several benefits of dealing with uncertainty (Reckhow 1996). Incorporating uncertainty in our analyses stimulates taking time to more fully understand the model's variables and their influence, often points to needed research, can provide context across applications when duly reported, and adds credibility to counter the "junk science" claim (Cleaves 1994). In effect, we are telling what we know and what we don't know – always a good idea. Including uncertainty as I have been discussing it also replaces the worst-case scenario. The worst-case scenario is good at getting us to think outside the rhomboid, but it has too low a probability of ever occurring to be taken very seriously (Suter, Barnthouse, and O'Neil 1987).

There is another issue I'd like to point out as well. Let's take what I believe is a thought-provoking example (Table 1), adapted from Evans (1982). Imagine for a moment a 100% accurate model. It predicts 'violations' of some temperature standard on 15 of 100 days in some given period. But we know our monitoring is less than perfect and can detect violations (or non-violations) only 95% of the time. The question is, what is the probability on any given day in that 100-day period that a violation appears to occur? Most people might initially say 95% of 15 -- 14 or so -- but that would be incorrect. The solution is achieved through the sometimes mysterious field of Baysian statistics. In reality we have 85 non-violations, but because of the inaccuracy of our measurements, only 81 appear to be non-violations and 4 appear to be violations. Similarly, we have 15 real violations, but 1 appears to be a non-violation and 14 are correctly characterized. So actually 22% appear to be violations. An open question persists:

how would the picture change if we acknowledge that we can never have a 100% accurate model?

One other problem has been a thorn in the side. Let's say that you run a Monte Carlo analysis as has been discussed, or arrived at uncertainty bounds by some other means and you characterize the model's uncertainty with a mean value plus or minus 1.5 degrees. Then you repeat the Monte Carlo simulation or simulate a "what-if" condition meant to describe a treatment option you are evaluating (let's say it is adding shade). This treatment reduces simulated temperatures by 6/10 degree (Figure 6). Some folks would look at this and say "no difference" but I believe they are wrong. Really the uncertainty bounds define the cloud on the whole model application and the proposed treatment shifts that whole cloud. This should be significant. We are going to need help from statisticians on this one.

Many other questions remain unanswered, too. Perhaps at the top of the list is how to treat the "margin of safety" criterion (NRC 2001). If our numerical standards allow for the standard to be exceeded a certain percentage of the time, can uncertainty analysis be used to tackle these problems? Borsuk, Stow, and Reckhow (2001) have some interesting ideas on the mechanics of the approach, but do not (and probably cannot) adequately address the trade-offs that must occur in the policy world or the issues involved with the Baysian statistics, as given in the Evans (1982) example.

## Conclusion

*No amount of sophistication is going to allay the fact that all your knowledge is about the past and all your decisions are about the future*. - Ian E. Wilson

Lack of an uncertainty analysis in water temperature modeling will not likely be seen as a limiting factor in analyses such as TMDL investigations (Boyd 2000). However, we need to learn what 'works' and what does not. We do not want to fall into the practice of incorporating uncertainty analysis simply to provide an illusion of scientific objectivity. Therefore, I recommend incorporating it in our analyses, but not without giving serious thought as to what uncertainty really means and how to improve the way we deal with it. Specifically, I recommend that we:

1. Acknowledge uncertainty in water quality modeling (NRC 2001). Use it as a part of identifying sensitive variables. Perform a Monte Carlo analysis even if it must be crude. However, keep uncertainty in perspective, as it is just one part of the process, and strive to use uncertainty to make better decisions, not simply quantify it (Conroy 1993).

2. Keep it simple stupid (KISS). Model builders need to beware of adding more detail to models to try to account for remaining 'error' because that 'error' may really be uncertainty in disguise and the added detail may introduce its own degree of uncertainty. Similarly, model users should beware 'over-fitting' models (Borsuk, Stow, and Reckhow 2001) to a small sample of measured data; small samples contain more uncertainty than large ones. I maintain that the best models are those that need minimal calibration.

3. Share your results. Communicate performance statistics so that we can begin to compare across applications (NCASI 2002). Report either the standard deviation of your estimates or the 95% confidence interval whenever possible, but beware being too enamored with statistics (Bunnell 1989); there is really nothing magic about any of these metrics. Report the distributions used in estimating the input variables so others can learn from your work. Present results graphically to policy makers so we can get experience with how this information is interpreted and used.

Beyond these technical recommendations, modelers need to accept more of a role as 'knowledge brokers' in our communications with policy makers (see Steel et al. 2001). We must take responsibility for communicating what uncertainty means and integrating it into decisions, all the while remembering that we will not assume the liability for dealing with the decisions made. That's a tough position to be in. And always remember that the uncertainties we have been discussing pale in comparison to the uncertainties in the realm of how the biological responds to thermal regimes.

## Acknowledgements

*I used to be uncertain but now I'm not so sure.* –Anonymous

I'd like to thank Geoffrey Poole for stimulating the desire to investigate this area, John Risley for providing the opportunity to communicate what I have learned, and many colleagues for letting me bounce ideas off of them, especially Brian Cade who helped me come to grips with some of the statistical concepts.

## Literature Cited

Bartholow, J.M. 2002. SSTEMP for Windows: The Stream Segment Temperature Model (Version 2.0). US Geological Survey computer model and documentation. Available on the Internet at http://www.mesc.usgs.gov/.

BioAnalysts, Inc. 2002 (draft). Application of new approaches to water quality temperature criteria: Chiwawa River Case Study. Report for Idaho Dept. of Env. Quality, Boise, ID. 37 p without appendices.

Borsuk, M.E., C.A. Stow, and K.H. Reckhow. 2001. Predicting the frequency of water quality standard violations: A probabilistic approach for TMDL development. Submitted to ES&T, August 23, 2001. Available on the Internet at http://www2.ncsu.edu/CIL/WRRI?Ken's_Page.html.

Boyd, J. 2000. The New Face of the Clean Water Act: A Critical Review of the EPA's Proposed TMDL Rules. Resources for the Future, Washington, D.C.. Discussion Paper 00-12, March 2000. 35 pp. Available on the Internet at http://www.rff.org/CFDOCS/disc_papers/PDF_files/0012.pdf.

Brown, L.C., and T.O. Barnwell. 1987. The enhanced stream water quality models QUAL2E and QUAL2E-UNCAS: Documentation and User Manual. EPA-600/3-87/007. USEPA Environmental Research Laboratory, Athens, GA.

Brown, L.C. ND. Modeling uncertainty – QUAL2E-UNCAS: A case study. Dept. of Civil and Environmental Engineering, Tufts University, Medford, MA. Available on the Internet at http://www.wef.org/pdffiles/TMDL/Brown.pdf. 6 p.

Bunnell, F.L. 1989. Alchemy and uncertainty: What good are models? General Technical Report PNW-GTR-232. Portland, OR, U.S. Dept. of Agriculture, Forest Service, Pacific Northwest Research Station. 27 p.

Cleaves, D.A. 1994. Assessing uncertainty in expert judgments about natural resources. General Technical Report SO-110. New Orleans, LA: USDA Forest Service, Southern Forest Experiment Station. 17 p.

Conroy, M.J. 1993. The use of models in natural resource management: Prediction, not prescription. Trans. 58th N.A. Wildl. & Natur. Resour. Conf. 509-519.

Essig, D. 1998. The dilemma of applying uniform temperature criteria in a diverse environment: An issue analysis. Idaho Division of Environmental Quality. Boise, ID. 29 pp.

Evans, J.S.B.T. 1982. Psychological pitfalls in forecasting. Futures. Pages 258-265.

Macaluso, Pat. 1983. Learning Simulation Techniques on a Microcomputer. Tab Books. Blue Ridge Summit, PA. 139 pp.

Macaluso, Pat. 1984. A Risky Business - An Introduction to Monte Carlo Venture Analysis. Byte Magazine. March, 1984. Pages 179-192.

NCASI. 2002. Excerpt from NCASI's comments to EPA on their draft Regional Temperature Guidance. Sent to me by Paul Wiegand, Regional Manager NCASI West Coast Regional Center, P.O. Box 458Corvallis, Oregon 97339, Phone: (541) 752-8801, email: PWiegand@wcrc-ncasi.org

National Research Council. 2001. Assessing the TMDL Approach to Water Quality Management. Committee to Assess the Scientific Basis of the Total Maximum Daily Load Approach to Water Pollution Reduction. Water Science and Technology Board. Division on Earth and Life Studies. National Academy Press, Washington, D.C. Available on the Internet at http://books.nap.edu/html/tmdl/.

Reckhow, K.H. 1994. Importance of scientific uncertainty in decision making. Environmental Management 18(2):161-166.

Reckhow, K.H. 1996. Uncertainty in water quality assessment. UNC Water Resources Institute. (Working Draft). Available over the Internet at http://www2.ncsu.edu/CIL/WRRI/Ken's_page.html. 11 p.

Rounds, S.A., T.M. Wood, and D.L. Lynch. 1998. Modeling discharge, temperature, and water quality in the Tualatin River, Oregon, with CE-QUAL-W2. Open File Report 98-186. U.S. Department of the Interior, U.S. Geological Survey. Portland, OR. 122 p.

ODEQ. 1999. HeatSource methodology Review: reach analysis of stream and river temperature dynamics. Oregon Dept. of Environmental Quality, Portland, OR. 13 pp.

Sullivan, K., J. Tooley, K. Doughty, J.E. Caldwell, and P. Knudsen. 1990. Evaluation of prediction models and characterization of stream temperature regimes in Washington. Timber/Fish/Wildlife Rep. No. TFW-WQ3-90-006. Washington Department of Natural Resources, Olympia, Washington. 224 pp.

Steel, B., D. Lach, P. List, and B. Shindler. 2001. The role of scientists in the natural resource and environmental policy process: A comparison of Canadian and American publics. J. Environmental Systems, 28(2):133-155.

Suter, G.W.II., L.W. Barnthouse, and R.V. O'Neill. 1987. Treatment of risk in environmental impact assessment. Environmental Management, 11(3):295-303.

Theurer, Fred D., Voos, Kenneth A., and Miller, William J. 1984 Instream Water Temperature Model. Instream Flow Inf. Pap. 16 Coop. Instream Flow and Aquatic System Group, U.S. Fish & Wildlife Service. Fort Collins, Colorado, approx. 200 pp.

**Tables**

Table 1. Example of Baysian approach to calculating probabilities of detecting water quality violations (adapted from Evans 1982).

| **Reality** | **Appear like non-violations** | **Appear like violations** |
|---|---|---|
| Non-violations = 85 | .95 x 85 = 81 | .05 x 85 = 4 |
| Violations = 15 | .05 x 15 = 1 | .95 x 15 = 14 |
| Total = 100 | 82 | 18 |
| Or 18/82 = 22% appear to be violations | | |

**Figures**

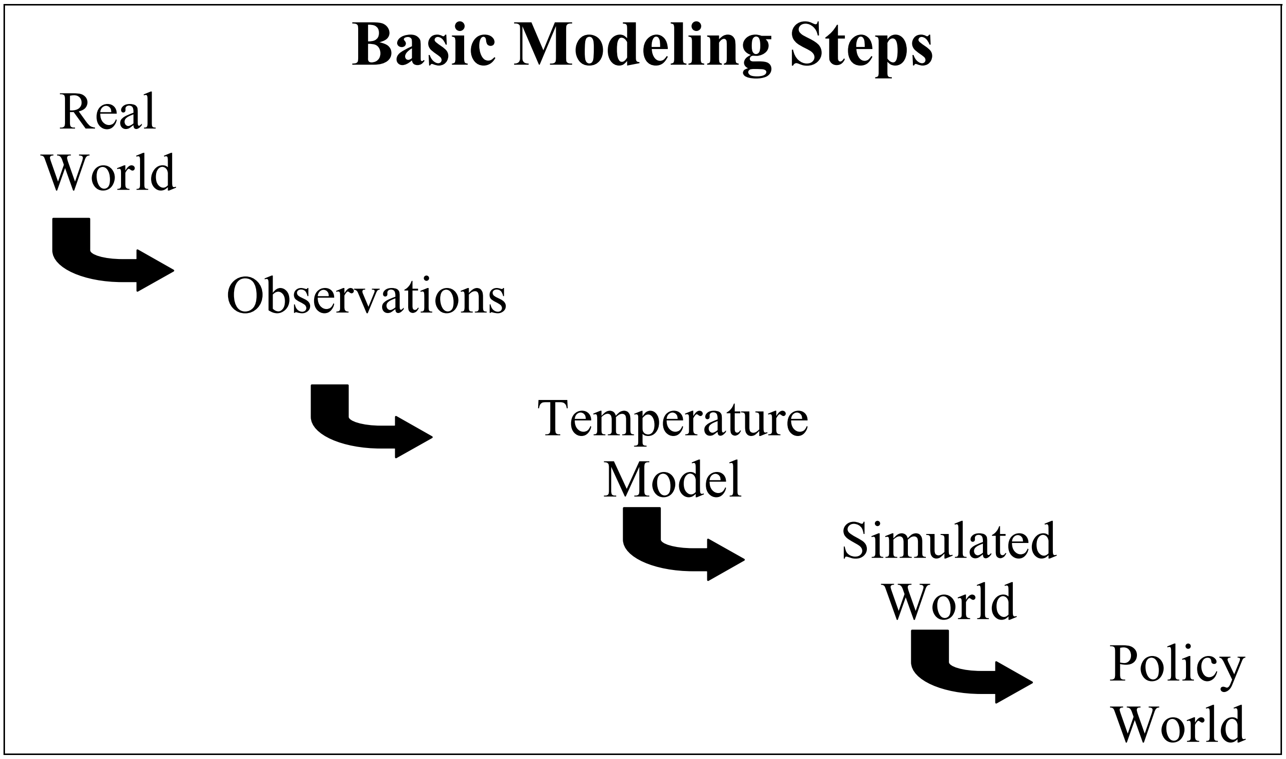


Figure 1. Important steps in the modeling process, from the real world to the policy world.

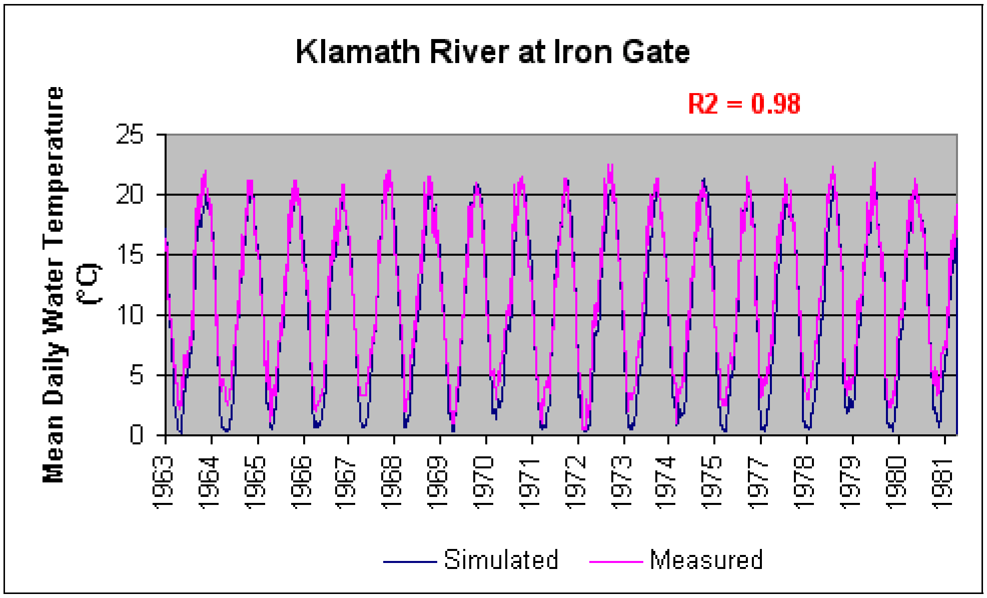


Figure 2. Sample goodness-of-fit plot indicating excellent fit of the model to the measured data.

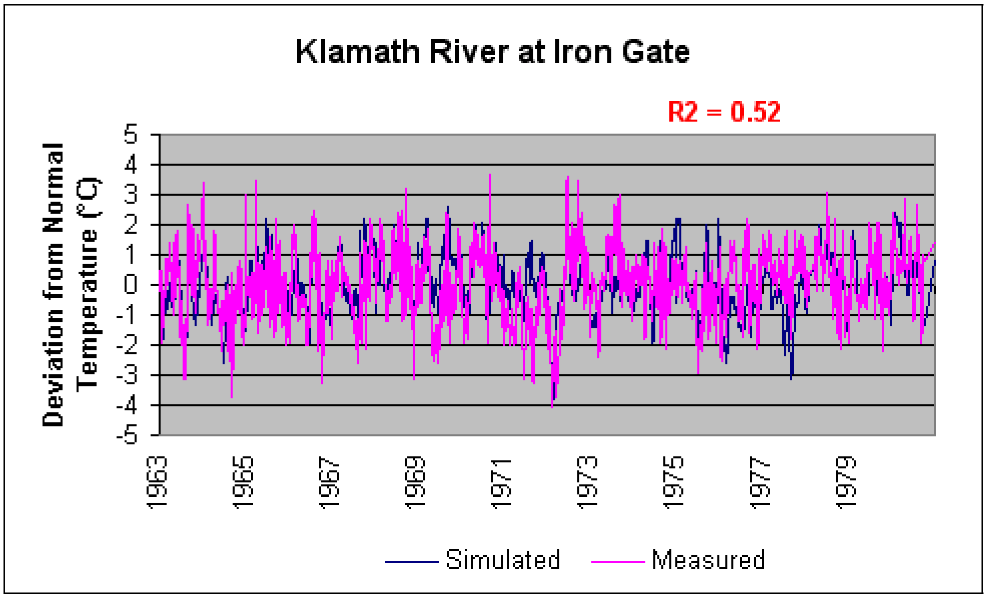
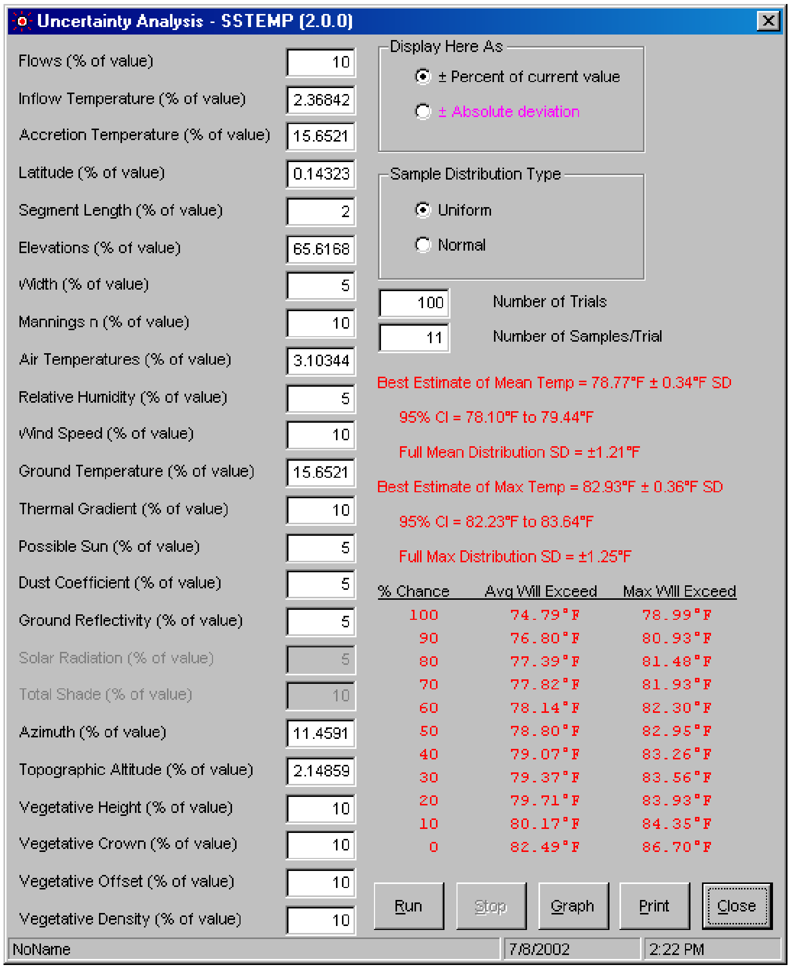


Figure 3. Same data as in Figure 3, but with autocorrelation removed.



Uncertainty Analysis - SSTEMP (2.0.0)

| | |
|---|---|
| Flows (% of value) | 10 |
| Inflow Temperature (% of value) | 2.36842 |
| Accretion Temperature (% of value) | 15.6521 |
| Latitude (% of value) | 0.14323 |
| Segment Length (% of value) | 2 |
| Elevations (% of value) | 65.6168 |
| Width (% of value) | 5 |
| Mannings n (% of value) | 10 |
| Air Temperatures (% of value) | 3.10344 |
| Relative Humidity (% of value) | 5 |
| Wind Speed (% of value) | 10 |
| Ground Temperature (% of value) | 15.6521 |
| Thermal Gradient (% of value) | 10 |
| Possible Sun (% of value) | 5 |
| Dust Coefficient (% of value) | 5 |
| Ground Reflectivity (% of value) | 5 |
| Solar Radiation (% of value) | 5 |
| Total Shade (% of value) | 10 |
| Azimuth (% of value) | 11.4591 |
| Topographic Altitude (% of value) | 2.14859 |
| Vegetative Height (% of value) | 10 |
| Vegetative Crown (% of value) | 10 |
| Vegetative Offset (% of value) | 10 |
| Vegetative Density (% of value) | 10 |

Display Here As

● ± Percent of current value

○ ± Absolute deviation

Sample Distribution Type

● Uniform

○ Normal

100 Number of Trials

11 Number of Samples/Trial

Best Estimate of Mean Temp = 78.77°F ± 0.34°F SD

95% CI = 78.10°F to 79.44°F

Full Mean Distribution SD = ±1.21°F

Best Estimate of Max Temp = 82.93°F ± 0.36°F SD

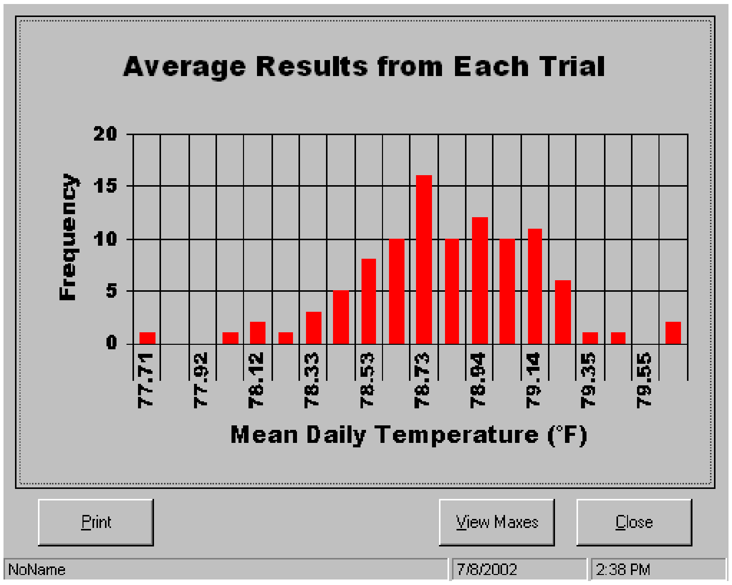
95% CI = 82.23°F to 83.64°F

Full Max Distribution SD = ±1.25°F

```
% Chance   Avg Will Exceed   Max Will Exceed
   100        74.79°F           78.99°F
    90        76.80°F           80.93°F
    80        77.39°F           81.48°F
    70        77.82°F           81.93°F
    60        78.14°F           82.30°F
    50        78.80°F           82.95°F
    40        79.07°F           83.26°F
    30        79.37°F           83.56°F
    20        79.71°F           83.93°F
    10        80.17°F           84.35°F
     0        82.49°F           86.70°F
```

NoName | 7/8/2002 | 2:22 PM

Figure 4. Adding input variable deviations to SSTEMP and text results.



NoName | 7/8/2002 | 2:38 PM

Figure 5. Results from 100 trials of SSTEMP uncertainty analysis.

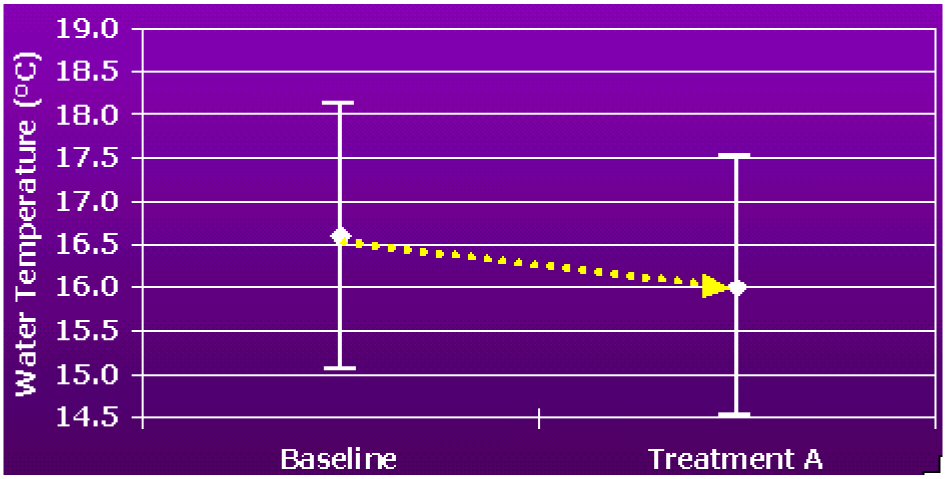


Figure 6.  Example of treatment effectiveness uncertainty.