# *Introduction to Environmental Modeling 1st edition Gray Solutions Manual*

SKU: 9781107571693-SOLUTIONS

# 2 Thoughts on Use of Data

Each year, in one of the first meetings of participants in the introductory course on environmental modeling, students were asked, "What would you do if an environmental accident occurred that caused a pollution event?" Almost without exception, the confident answer given was, "I would collect data." This answer must be scrutinized: What data? How often? Where? What does the data mean? How should the data be used? Answers to these questions are much more difficult. The purpose of this chapter is to begin to encourage consideration of data and its uses. Data is physical, and students welcome this aspect. Putting the data into a context that aids in understanding of mechanisms naturally requires some knowledge of mechanisms. The fact that modeling requires use of data in conjunction with an understanding of not only the questions that one is trying to answer but also the factors that impact changes in data is introduced here. This chapter is not concerned with statistical analysis of data but is meant to identify some pitfalls and considerations when data is used in conjunction with deterministic models.

## 2.1 Introduction

## 2.2 On Numbers in Elementary Education

## 2.3 What's the Answer?

**Question 2.1:** What is the answer to this question?

Of course in the absence of all context, this question is absurd! This question is raised as a counterpoint to the types of questions that students overwhelmingly encounter during their educational career: questions that are perfectly posed such that all necessary information



Introduction to Environmental Modeling 1st edition Gray Solutions Manual

is provided, no excess information is given, and a precise and unique solution can be obtained making use of a provided tool. In fact, such "perfect" question are often encountered in environmental modeling courses. These "perfect" questions are rarely encountered in fact. Modelers have to determine what questions need to be answered and how to answer them. Yet, expectations are that "perfect" questions are closer to the norm than the "absurd" question posed here. Additional scoffing and ridicule for this question appear in the text.

---

**Question 2.2:** Given the following data for the first realization of a process:

$$(n, y) = (1, 10)$$

What will be the value of the second realization? In other words, what is the value of $y$ for the pair $(2, y)$?

---

After the discomfort typically induced by the absurdity of Question 2.1, this question may provide a sense of relief. The reflexive reaction to correlate data kicks in. However, in the absence of any information about an operative process, the next data point can be anywhere from $-\infty$ to $+\infty$. At least by restricting the discussion to numerical answers in the text - precluding a color, a name, a vegetable, a country, anything - the scope of consideration has been limited somewhat. It is a somewhat rare student who, on seeing this answer does not think that the next data value will be a number somewhere in the vicinity of "10." This tendency has to be resisted, and the notion of being "in the vicinity" can be discussed.

---

**Question 2.3:** Given the following data for the first two realizations of a process:

$$\{n, y\} = \{(1, 70), (2, 72)\}$$

What will be the value of the third realization? In other words, what is the value of $y$ for the pair $(3, y)$? Also, what would the value of the second data point have been if the measurement were taken between the two realizations indicated?

---

This continues along the path from the previous two questions. Most students would tend to think that the answer is "74", although there will be some reluctance to speak up because of the lessons from the text. This question can lead to the question, "How many data points does one need to be confident with a projected 'next' answer to a problem?" From Question 2.1, no data is not satisfactory; from Question 2.2, one data point is unsatisfactory; at some point the desire to infer an answer may overwhelm the need to be objective. The factor that can assist in determining when such an inference is justified is some knowledge of the process that produces a data point value. Without any knowledge, the quality of the selected next value is poor, even if it turns out to be the correct value!

Introduction to Environmental Modeling 1st edition Gray Solutions Manual

**Table 2.1** What's the value of $y(4)$?

| n | y |
|---|---|
| 1 | 1 |
| 2 | 10 |
| 3 | 100 |
| 4 | ? |

The second part of this question deals with interpolation. The answer that calls out like a Siren is "71"! When knowledge of a process is present, interpolation may be safer than extrapolation. In the absence of knowledge of the process, interpolation is no better than extrapolation.

Useful comments might be made about what "no knowledge" of a process means. Even small bits of qualitative information can be helpful in reducing the range of quantitative answers to a question.

$$\log_{10} y = n - 1 \tag{2.2}$$

## 2.4 Given the Process, What's the Answer?

## 2.5 Given the Answer, What's the Process?

**Question 2.4:** List some "laws" that you have experienced in your engineering and science background that might more properly be thought of as approximations or correlations. What are the limitations on each approximation?

Candidate answers will vary depending on students' backgrounds. Possibilities include the ideal gas law, Moore's law, Fourier's law, Birch's law, Newton's laws of mechanics, Darcy's law, Bernoulli's equation, Archie's law, Law of the wall, Toricelli's law, Fick's law, for example. Each law has its own shortcomings and limitations. Answers to this question should include the concept that it is as important to know the limitations of any law used to describe a system as it is to apply the law.

Introduction to Environmental Modeling 1st edition Gray Solutions Manual

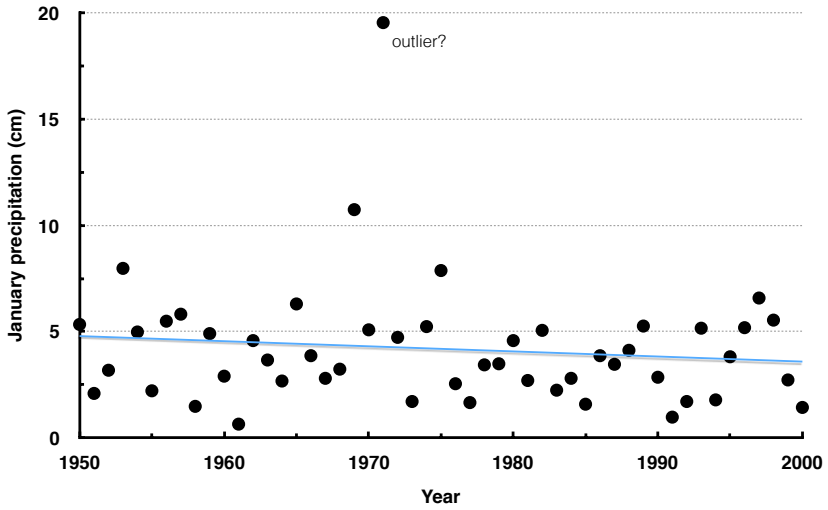



Fig. 2.1 Plot of available data for January rainfall in Grangeville, Idaho from 1950 to 2000.

## 2.6 Given an Answer, Is It Useful?

### 2.6.1 Given the process, is the proposed answer useful?

**Question 2.5:** Think of an example from your engineering and science experience in which an answer or result produced by a designed process was unsatisfactory because the designed process was deficient. How could that process have been modified or improved to produce a better outcome?

This answer will vary depending on the process considered. Elements of an answer could include budget, ability of personnel to execute the task, quality of the equipment, computer resources, collect more data in time and/or space, understand the process before collecting data, and checking resources for insights from other studies. Depending on the design process, other elements of the study might be reconsidered.

**Question 2.6:** Identify a naturally occurring environmental phenomenon or event in

Introduction to Environmental Modeling 1st edition Gray Solutions Manual

which an answer or prediction from a model was not satisfactory because the description of the process employed in the model was inadequate.

Students might go online or check the library to find examples. Possible scenarios include collapse of cod stocks; excessive groundwater withdrawal; models of hurricane threats to the Gulf coast; earthquake prediction. Since success is touted more than failure, the role of models in failed management of resources tends to be overlooked in the literature.

### 2.6.2 Given the data, is the answer useful?

**Question 2.7:** Describe an engineering or science experiment and the type of data it produces. In what situations might the experiment produce data that leads to a more useful answer? Less useful?

This question requires reflection by the student on an experiment of interest. Issues of time scale, length scale, uncertainty, and process description might enter into the description of situations.

**Question 2.8:** Consider the following limited data. The average summer temperature in the lower 48 United States was 23.56°C in 2012, 23.61°C in 2011, and 23.67°C in 1936. The average temperature in May, 2015 in Bergen, Norway was 8.72°C, the coldest May since 1979 and well below the recent 50 year average of 10.50°C. Discuss this data in light of the efforts to study and predict global climate change. What are the shortcomings, utility, and value of this data? What additional data might be useful?

This data has been "cherry picked." The fact that it is for limited regions and over limited time ranges means that it is of limited utility by itself, but it could be useful in conjunction with other data and knowledge of processes concerning climate. The challenge is to determine how much data is needed as a function of time and space to be able to confirm or reject a prediction from a model that, itself, is able to account for only some of the operative processes.

Introduction to Environmental Modeling 1st edition Gray Solutions Manual

## 2.7 Problems

These problems relate to what one might infer from data. Thorough analysis of the data requires statistical analysis. However, the prime focus of this text is on deterministic modeling. Thus, the problems examine data for the purpose of indicating that one must be careful in declaring what data indicates. The fact that averaging diminishes the variability should be emphasized in discussion. This serves to hide large and small values, that may be important for knowing system behavior, with the benefit of making longer term trends clearer. An optimal approach to advanced modeling combines statistical methods in examination of data and in specification of system properties with mechanistic descriptions of operative processes.

## Problems

**2.1** Based on the data plotted in Fig 2.1, can one reliably predict that average precipitation over the first ten years of the twenty-first century in Grangeville will decrease or increase? Justify your answer.

Based on the trend line, it seems that the average precipitation is likely to decrease over the first decade of the twenty-first century. The data set has been replotted for the years from 1950-2000, with the outlier removed, in Fig 2.2. The data set for the years 2000-2010 has been plotted in Fig 2.3. Finally, the data set for the full period from 1950-2014 appears in Fig 2.4. These plots all consist of monthly averages for January. In all cases, the trend is for the rainfall to decrease on average. However, within one standard deviation of the data, the average could be increasing or decreasing. Thus, although the trend from the 1950-2000 data was continued into the twenty-first century, this does not mean that each year will produce lower precipitation than the previous year. The data set, which spans about 60 years, is not adequate to predict what will happen in the coming 60 years. Using the data set implies nothing about the mechanisms involved. We do know that the precipitation measurement will be non-negative. To get an improved estimate of what might happen in the future, the data would have to be considered in the context of the processes involved, both on the local scale in Grangeville, and as a result of larger scale global processes.

**2.2** Look up the data for the years from 2001-2010 and determine the average precipitation in Grangeville, ID. Is it less than the average for the preceding 50 years? For

Introduction to Environmental Modeling 1st edition Gray Solutions Manual

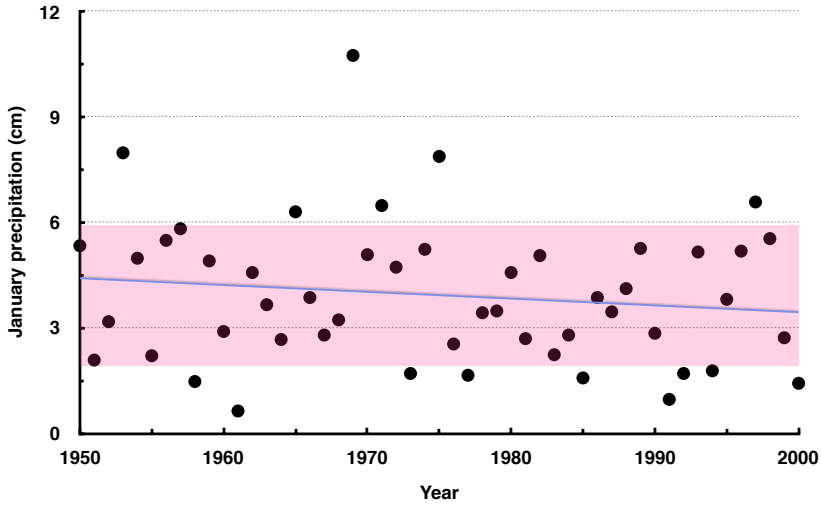



Fig. 2.2 Plot of available precipitation data for January rainfall in Grangeville, Idaho from 1950 to 2000 with linear fit. Pink band represents the standard deviation around the mean.

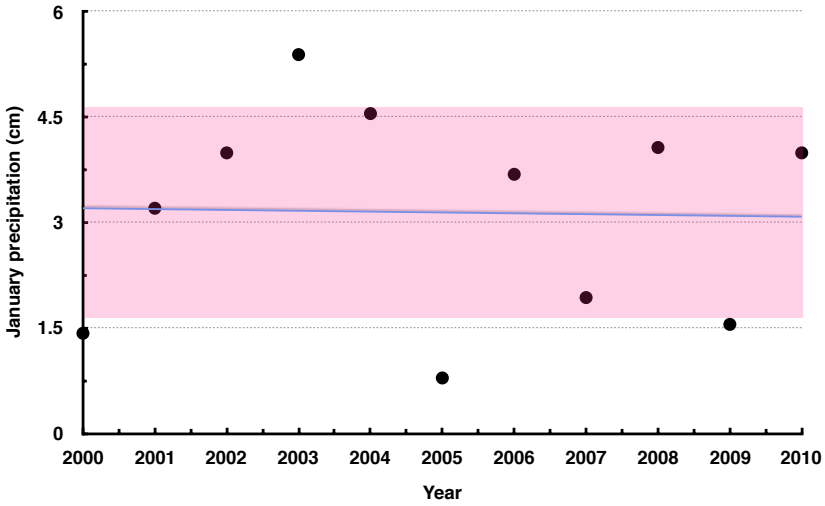



Fig. 2.3 Plot of available precipitation data for January rainfall in Grangeville, Idaho from 2000 to 2010 with linear fit. Pink band represents the standard deviation around the mean.

the preceding 10 years? What does this tell you about the use of the rainfall data as a predictor?

This data is available at
`http://weather-warehouse.com/WeatherHistory/PastWeatherData_Grangeville_Grangeville_ID_January.html`
For the years 2000–2010, the average precipitation in January was 3.140 cm.

Introduction to Environmental Modeling 1st edition Gray Solutions Manual

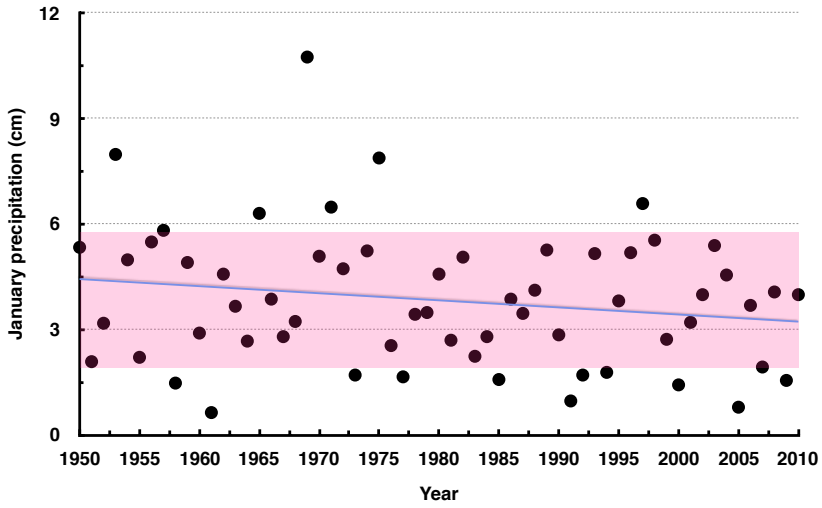



Fig. 2.4 Plot of available precipitation data for January rainfall in Grangeville, Idaho from 1950 to 2010 with linear fit. Pink band represents the standard deviation around the mean.

For 1950–2000, the mean precipitation in January was 3.927 cm.
For 1950–2010, the mean precipitation in January was 3.826 cm.
For 1990–2000, the mean precipitation in January was 3.427 cm.
The average values of the precipitation show a consistent trend. Although the January values each year tend to vary, the ten year average between 1990–2000 and 2000-2010 show a small decline consistent with the larger trend. One cannot say that the next ten year period will exhibit a corresponding decrease. Averaging of the data does smooth out the trend.

For comparison, a plot of the $\log_{10}$ of the January precipitation data appears in Fig 2.5. Because precipitation is not negative, a plot of the log of the precipitation may better correspond to a normal distribution. The mean of the log of the precipitation is 0.5219 which corresponds to a precipitation amount of 3.326 cm. This result can be used to point out the importance of how data is handled.

**2.3** Consider again the data for the years 2001-2010 found for Problem 2.2.

(a) Double the value of the data found for 2005 and determine the new average precipitation in Grangeville, ID. How different is it from the average found using the correct data? How "wrong" do you think the data can be and still produce a "correct" average?

With the data value doubled, the average for the ten year period from 2000–2010 increases from 3. 140 cm to 3.212 cm. This is an increase of 0.07 cm. Thus an error that might double the measurement in 2005 produces an error of a little more than

Introduction to Environmental Modeling 1st edition Gray Solutions Manual

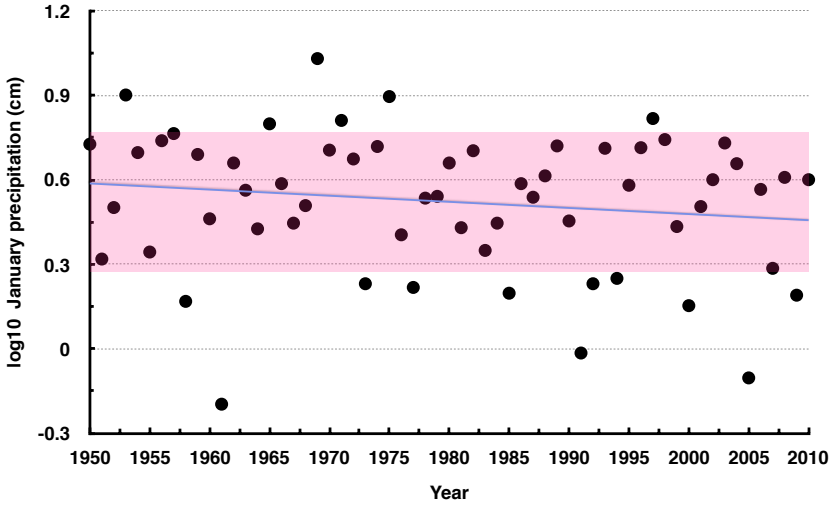



Fig. 2.5 Plot of $\log_{10}$ of precipitation data in cm for January rainfall in Grangeville, Idaho from 1950 to 2010 with linear fit. Pink band represents the standard deviation around the mean.

2% for the average over the 10 year period. This result raises the question of how the monthly rates were determined. Although it is a small town, Grangeville is not a point. Rain gages at different locations can give different readings. If the reading in the rain gage for 2003 were half what a "correct" reading would have been, the average rainfall from 2000–2010 would actually be 3.630 cm. This would correspond to an increase in the actual rainfall for the decade over the previous decade. It is clear that knowledge about the accuracy of the rain gage, of how the data was collected, and of variability among different gages in the area would all be important elements to determining if the January averages are correct.

It might be informative or serve as an alternative exercise to look at running averages of rainfall using 5 year windows, or some other measure, to demonstrate the smoothing effect of averaging.

(b) Given the 10 values of precipitation in Grangeville, ID for 2001-2010, find the 8 that give an average closest to the 10 year average. Given the data for 50 years, how many of these values do you think would be needed to give a realistic predictor of the average yearly precipitation? What does this tell you about using parts of data sets to make determinations?

From Fig 2.3, the values that are farthest from the mean are the precipitation rates in 2003 and 2005. If these are removed from the data set, the average precipitation is reduced from 3.140 cm using all ten values to 3.129 cm based on the eight values. This is not a recommended approach as data has been skipped because of its

Introduction to Environmental Modeling 1st edition Gray Solutions Manual

value. An alternative would be to examine the average of the even years from 2000 to 2010 (3.615 cm) vs. the average of the odd years in this interval (2.57048 cm). These changes in average, along with the fact that the standard deviation is about 1.5 cm, suggest that use of 5 or 6 data points does not give a stable average. When using the even years from 1950–2000, the average precipitation is 3.751 cm while the odd years in this period average to 4.109 cm. These differences suggest a number interesting thoughts: 1) the data set of 50 years is not a data set for a long period of time; 2) one should not suggest that the likelihood of heavier than average precipitation in the future is greater in an odd year than in an even year; 3) understanding of natural variability in the data is important in being able to draw conclusions from the data. In this case where the standard deviation of the precipitation rate is approximately half the rate, variability is important in making plans that would manage this resource.

**2.4** An alternative to Eqn (2.2) for modeling the data in Table 2.1 is

$$\log_{10} y = n - 1 + A \log_{10}[|\cos(2n\pi)|] \,, \tag{2.5}$$

where $A$ is a constant. Suppose that $n$ is some measure of time such that the equal intervals in $n$ in the table indicate that the data has been collected at equal time intervals.

(a) Does Eqn (2.2) or Eqn (2.5) provide a better fit of the data in Table 2.1? What value of $A$ provides the best fit?

Both equations fit the data perfectly. For the way the problem is posed, $A$ can take on any value while still providing a perfect fit of the data.

(b) If you are not informed about the physical processes occurring, discuss why you might prefer one of the equations as opposed to the other as a description of a system.

In the way the problem is posed, with $n$ being a measure of time, non-integer values of $n$ will be appropriate. With non-zero values, the term involving the cosine comes into play. Because the absolute value of the cosine ranges from 0 to 1, the logarithm will always be negative, and will have a magnitude of $\infty$ whenever $n = m + 1/4$ or $n = m + 3/4$ and $m$ is an integer. Without knowing about the process involved, yet knowing that there is a process, Eqn (2.2) is a safer pick. It presumes monotonic increase in $y$ while Eqn (2.5) is highly oscillatory. It could well be that neither selection is a good selection, as illustrated by the story of the family reunion in the text.

(c) If you are given access to the experiment that produced the data in Table 2.1, how

Introduction to Environmental Modeling 1st edition Gray Solutions Manual

would you go about determining if Eqn (2.2) or Eqn (2.5) provides a better description of the system or if some other equation might be superior?

A quick text for comparison of the two candidate equations would be to run the experiment at a time corresponding to $n = 1/4$. At this value, Eqn (2.2) provides a finite value of $y$ while Eqn (2.5) provides a large value (except when $A = 0$ and the equation reduces to Eqn (2.2)). To further study equations for fitting the data, one could perform experiments with non-integer values of $n$ to fill in the data set and then search for a fit that seems to fit values of the data and intermediate values. Of course, high frequency variability of the data will not be observed if the increment of data collection is too large.

**2.5** The fact that lunar tidal processes operate on a period of 12.4 hours has some implications on data collection. Discuss how your knowledge of this process impacts how you might design experiments regarding flow processes, presence of fish species, or contamination in a near shore region.

Students might draw a sine wave with a period of T and see how this wave is represented if data is collected at intervals of $3T/2$, $T$, $0.5T$, $0.25T$, and $0.1T$. Depending on the location of the first data point, various sine waves will be suggested. For example, if one uses data at $\{0, 3T/2, 3T, 9T/2\}$, one gets no indication of the tidal process. However, if data collected at $\{T/4, 7T/4, 13T/4, 19T/4\}$ suggests a sine wave with a period of $3T$ (which is longer than the actual period). Collection of data at an interval of $0.1T$ suggests a sine wave with the period T. Thus, data should be collected so that values are collected at some time increment that is a fraction of the wave period. This will ensure that information between high and low tide will be represented properly. In a typical situation, one might not know the actual period. Thus, the data collection interval should be decreased until a stable period emerges. Note also that in additional to the principal lunar tide, denoted as $M_2$ with a period of 12.42 hrs, other tidal components will impact measurements (e.g., $S_2$ with period of 12.0 hrs, $N_2$ with period of 12.66 hrs, $K_1$ with period of 23.93 hrs, $O_1$ with period of 25.82 hrs, $P_1$ with period of 24.07 hrs.), and some smaller components that have longer periods. The $M_2$ component is the largest tidal component.

Introduction to Environmental Modeling 1st edition Gray Solutions Manual