



Prof. Yoram Rubin, Ph.D  
Department of Civil and Environmental Engineering  
UC Berkeley  
627 Davis Hall  
Berkeley, California 94720-1710  
Tel. 510-642-2282  
e-mail: rubin@ce.berkeley.edu

October 7<sup>th</sup>, 2011

**Ms. Lauri Kemper, PE**  
**Assistant Executive Officer**  
**California Regional Water Quality Control Board**  
**Lahontan Region**

**Re: Peer Review of PG&E's Chromium Background Study Report, Hinkley Compressor Station**

This review provides my opinions on several questions related to the documents provided to me. The review is organized following the sequence of questions raised in the Scientific Peer Review Request (Sections 1-4). Additional comments of a more general nature are provided in Section 5. If I missed or misinterpreted any information, I would be glad to be informed about it.

**1. Comments on quality of spatial sampling of background chromium**

The first issue raised in the "Scientific Peer Review Request" concerns the large number of wells installed (and measurements taken) in the vicinity of well BGS-04. Looking at Figures 4-1 and 4-2, it is obvious to me that there are many measurements collected all over the site, and altogether they form a good basis for analysis and for making predictions. The challenge of course is how to analyze the data and how to use it for predictions. Specifically, there is a need to apply analysis that would take into consideration that uneven spatial distribution of the measurement locations (i.e., the sampling wells). Without taking this into account, the concentrations at a particular area (e.g., BGS-04) could be assigned a disproportionately large weight. If many or all the wells around BGS-04 sample a particularly high concentration area, the high concentration in that area could pull the spatial average higher (creating a positive bias), leading to averages that are not representative of the site. It could also happen that they all sample small values, and that would create a negative bias. This is known in geostatistics as the clustering effect. The clustering effect could be removed through declustering. It does not appear that declustering was applied to the data. To summarize, the uneven distribution of wells could lead to bias. There are known techniques that could handle the clustering effect, but none was carried out, to my understanding.

Additional comments:

1. The Background Study mentions on page 1-4 that “To compensate for the lack of discrete-depth-samples, PG&E proposed to expand the background study well network”. In response to that statement, this approach cannot work unless the concentration field is stationary and statistically isotropic, which cannot be the case. So, expanding the area being sampled cannot compensate for the lack of discrete-depth samples.
2. Table 3-1 indicates that several of the wells are screened over the upper (floodplain) and lower (regional) aquifer. From my understanding of the sampling procedures (Section 3.2), the concentrations represent (flux-) averages over the entire screen. This could lead to ambiguity as to what the concentration averages actually represent (i.e., which geological unit?). Furthermore, it could also lead to bias: it may be that a well that is screened over the two aquifers would mix clean water from one unit with contaminated water from the other unit, which would lead to biases when trying to assign the measured value to a particular aquifer and to biases in assessing the average concentrations. This ambiguity could be removed, to a large degree, through appropriate modeling, but to my understanding this has not been done.
3. Spatial averages are of little predictive value in the case of non-stationary variables such as the concentration. The population sample mixes measurements taken upstream (potentially low values) and downstream (potentially larger values) of the compression area. There also appears to be a trend of the concentrations increasing from east to west. All this could lead to biases. A physically-based analysis could take the trends in the concentration into account and provide better predictions.

**2. Comments on quality of temporal sampling of background chromium.**

The procedure used to account for gaps in the temporal sampling is described as follows (Scientific Peer Review Request, Attachment 2)

**2. Quality of temporal sampling of background chromium**

The 2007 Background Study Report acknowledges that the expansion of the well network after the second sampling event has the potential to introduce bias into the overall summary statistics due to the temporally unbalanced nature of the data set (i.e., four quarters of data are not available for all wells). To address this bias, the arithmetic average value of Cr(VI) and Cr(T) concentrations from each well were used in the statistical analysis. Therefore, each well is represented by one arithmetic mean result instead of by the actual number of samples taken at that well. See the 2007 Background Study Report, pages 5-5 through 5-7, and page 7-1.

I find this approach lacking in several respects, and I would recommend against it. My reasons are as follows. Averaging is known to alter the statistical nature of the variables being averaged. The primary effect is reducing variability. The consequence of that is that the averaged variables provide a “smoother” version of reality, and as a result the high and low values are averaged out. The elimination of high values of the concentration from consideration is obviously of concern in the context of this study because it would lead to biased estimates.

Appendix I of the Background Study Report refers to this issues and mentions the “..dampening the effect of the most elevated values in the sample set by averaging those results with lower results from other sampling” (page 7-1). I cannot see why dampening would be a desired outcome. To explain this issue consider the following example: if you are searching for gold, you will not average gold concentrations from your soil samples, because that one sample with very high concentrations could be very important in telling you where to dig. Similarly, the samples with high concentrations could indicate the presence of high-concentration areas and should not be averaged out.

There is another problem with averaging of measurements that is related to the test of statistical normality (discussed further in Section 3 below). Statistical tests are generally performed (unless stated otherwise) based on statistically homogenous populations (population samples), meaning that all samples in the population sample are drawn from (or representative of) the same underlying distribution. In many cases, the samples are assumed to be independent and identically distributed (what’s known in the statistical literature as i.i.d). The assumption of homogeneity is a key element of statistical inference. Averaging as done in the Background Study is inconsistent with this requirement, because the averaged concentrations and the non-averaged concentrations do not belong in the same underlying statistical distribution. I will discuss this issue further in Section 3, but in brief summary, the mixing of variables from different distributions violates one of the assumptions used to construct the Shapiro-Wilk test. The consequences of this violation were not evaluated and so cannot be ignored.

### **3. Comments on the assumption of statistical normality.**

The normal distribution is a favorite model selection in applications because of its simplicity: one needs to infer only 2 parameters (the mean and variance) to be able to define the entire distribution, which could then be used for making predictions and associating them with confidence intervals. Given that in groundwater applications there is not a lot of data to begin with, and that inference of multi-parameter models is a challenge, there’s no wonder why one would want to adopt the normal model, as was done in the background study.

In order to test whether or not a normal model is acceptable, the background study elected to use the formalism of hypothesis testing. The underlying theory is documented in many textbooks. The approach is to state a null hypothesis (in this case, that the concentrations are normally distributed) and then to apply a test that would indicate whether this assumption could be rejected or not. A fundamental tenet of hypothesis

testing is that the test can only determine whether there's enough evidence to reject the null hypothesis. Hypothesis testing does not provide conclusive evidence that the null hypothesis is the right one. It can only determine whether or not there's enough evidence to reject it. Based on this, the statement made in Appendix I that "the probabilities (p-values) from the Shapiro-Wilk test (W test) provide evidence about whether the background total and hexavalent chromium concentrations are normally or log-normally<sup>1</sup> distributed" is very doubtful. The test does not provide such evidence, its power is only to state whether there's enough evidence to reject the assumption of normality.

Not having enough evidence to reject the null hypothesis (normality) does not mean that the normal model is the best one. It also does not mean that other evidence cannot be used. To use an analogy, not finding conclusive evidence with fingerprints does not mean that DNA samples cannot be used and shed a different light. In the case of the normal model assumption, it should be noted that the concentration is by definition non-negative, and hence non-normal by definition (exceptions can be made but I am not sure they are applicable here). There is evidence for asymmetry in Table 6.1 where differences between the mean and median of the distribution are shown to exist: in normal distributions these values should be equal (or at least very close to each other). Hence, there are indications against the assumption of normality.

The practice of hypothesis testing brings another issue to the surface. In hypothesis testing, the common thinking is that the null hypothesis should be a "safe" assumption, meaning an assumption that would not lead to damage if it is not rejected. This is because it is difficult to reject the null hypothesis: it is rejected only in the face of overwhelming evidence against it. Let me explain this with an example from the criminal law. I am not a jurist, but this example is commonly used and I think I understand it pretty well. The point is that legally a person is assumed innocent until proven guilty. So the null hypothesis in the legal system is that the person is innocent. The assumption of innocence is selected to be the safe assumption (null hypothesis) in most legal systems, and it will be rejected only in the face of overwhelming evidence to the contrary. How is that related to the Background Study? The question is whether the assumption of normality is the safe assumption and should it be used as the null hypothesis. In my opinion it is not a safe assumption because it could underestimate the probabilities of high concentrations. For example, a lognormal distribution has a longer "tail" and it assigns higher probabilities to the high concentrations, and so it could possibly be a safer assumption. This option and perhaps others need to be considered.

The quality of the sample population is obviously of primary consideration. Shapiro and Wilk (1965) assume that their samples are identically distributed. Section 2.2 in the Shapiro-Wilk paper states that "The objective is to derive a test for the hypothesis that this is a sample from a normal distribution with unknown mean  $\mu$  and unknown variance  $\sigma^2$ ." As discussed in Section 2, the sample population includes measured concentrations and averaged measured concentrations. Because averaging alters the statistical nature of the underlying distribution, the population sample appears to be inappropriate for this

---

<sup>1</sup> Shapiro and Wilk (1965) mention only the normal option, not lognormal. The log-normal option is a possibility after log-transformation of the measurements.

kind of test because differences in temporal averaging procedures (e.g., averaging over 2, or 3 or 4 measurements) will lead to different statistical distributions for the various samples within the population sample, in a violation of the requirements of the test. The consequences of such violation need to be analyzed, but in principle, inferences from such a hybrid sample population are not suitable for determining the nature of the underlying distribution.

The Background Study does not assume correlation between the concentration measurements. In other words, the measurements are assumed to be spatially-uncorrelated. This assumption, although not unreasonable for measurements with large distances in between, is not justified theoretically, and is particularly challenging for measurements at close proximity. It needs to be supported with evidence. I could not find such evidence in the study and I am concerned that the test is inconsistent with the underlying physics.

In another direction, the test of normality addresses the question of whether or not the population sample could be described as normally-distributed. It does not address the question of whether or not the normal model inferred from the population sample is a good model for prediction of regional or local averages of the concentration and its confidence intervals. More on that is provided in Section 5.

In light of this discussion, I believe that the outcome of the Shapiro-Wilk test is questionable. Additional comments on this matter are provided in Section 5.

#### **4. Comments on quality of groundwater modeling**

The groundwater model is discussed in Appendix B. Model calibration is discussed in Section B.1.4. Very little information is provided and whatever is given is not enough to confirm the adequacy of the calibration effort. Particular issues to consider are as follows:

1. The model was calibrated based on groundwater levels only. This raises several issues of concern:
  - a. Water levels alone cannot be used for calibrating the spatial distribution of the hydraulic conductivity because there is no unique relationship between water levels and conductivity. Without sound calibration of the hydraulic conductivity field and porosity, the groundwater model cannot be used to predict velocities, and concentrations.
  - b. No information is provided on the quality of the match between measured head and model-based predictions. It is important to remember in this context that even small errors in the predicted heads could lead to very large errors in the head gradients, and all that is related like velocities and concentrations.
  - c. Without reliable estimates for the hydraulic conductivity, the reliability of the water budget analysis cannot be established.
2. No attempt is reported to test the model against the concentration data. This could be a useful strategy to establish the credibility of the model. Methods for using concentration data are available (see Rubin, 2003 and Rubin et al., 2010).

3. No attempt to model spatial variability of the hydrologic parameters is reported. Assuming the hydraulic conductivity to be uniform within each of the hydrostratigraphic units would neglect the possible consequences of channeling effects that could be introduced by the "...interbedded gravels, sands, silts, and minor amounts of clay " (Section B.1.2). One possible consequence is that the channels could act as fast flow channels. Such channels would lead to faster downstream migration of chemicals.

My conclusion is that more work is needed in order to align the model calibration efforts with modern concepts on this topic. As discussed in Section 5, uncertainty quantification (UQ) should be an important part of the study. A groundwater model is the main vehicle for UQ. This line of thinking was not pursued here and no UQ that meets acceptable norms was carried out, to my understanding.

## 5. General comments

In Section 3 I addressed questions related to the normality test. Here I would like to provide additional perspective. The first point I would like to make is that, regardless of whether or not the Shapiro-Wilk test is applicable or not, there is a need to evaluate the predictive capabilities of the normal model, and that is a different issue altogether. In other words, even if one accepts that the population sample is normal (see Section 3 for discussion on the difficulties with this), this does not constitute a confirmation that the normal model could actually be used for predicting (at best) anything but the statistics of that population sample, until the predictive capability itself is tested. The main reason for that is the issue of ergodicity. For spatial averages to be representative, the population sample must be ergodic (see Rubin, 2003). That means that the population sample must cover all the possible states of the sampled system, and in the right proportions. If this condition is met, then the population sample would be sufficient for making inferences about spatial averages. For stationary problems, satisfying the condition of ergodicity requires extensive spatial sampling. How large the sampled domain needs to be? This can only be established through physically-based modeling of the aquifer, including modeling of the spatial variability of the hydraulic conductivity and the flow and transport fields related to the spatial variability model. The added complication here is that the concentration field is non-stationary. This could be compensated through physically-based stochastic modeling strategies (Rubin, 2003). Another strategy to evaluate the model's predictive capability is through cross-validation (Rubin, 2003).

Another issue to consider is the no-detect concentrations. Figures 5-4 and 5-5 and associated discussion indicate that locations where the concentrations were measured below the detect limits were assigned values equal to half the detection limit. This is speculative. It may be a good speculation, but it is still a speculation, nonetheless. The speculation is in considering and analyzing the concentration from the perspective of a spatially-uncorrelated variable rather than a spatially-correlated variable. The point is that if one adopts the spatial correlation perspective, the no-detects could be interpreted in different ways. For example, one could also speculate that the no-detects could be indications of fast-flow channels with very high concentrations further downstream

(Wilson and Rubin, 2002), or that the wells with no-detects were placed in low-conductivity areas with by-pass flow nearby.

At times one must resort to speculations when it comes to groundwater applications, but there is a need to establish their likelihood. What is needed here is to substantiate this speculation by evaluating it using a physically-based flow and transport model. Another important point is that including speculative values in the population sample used to test normality is not warranted. Without accounting for the uncertainty around this speculation, one cannot assign any confidence intervals to any prediction that is based on a population sample that includes these values. This adds further doubts to the value of the normality test (see Section 3 for additional discussion).

The next comment is with regard to uncertainty quantification (UQ). UQ is the idea that all sources of uncertainty must be accounted for when making predictions. It is known that the sources for uncertainty are spatial variability and data scarcity, and the challenge is how to quantify that uncertainty. To be specific with regard to the analysis carried out in the Background Study, we would want to model the model uncertainty (in other words, how likely or unlikely is the normal model and alternative models?) and the parameter uncertainty (in other words, what is the uncertainty associated with the parameters of the normal model?). UQ is a fundamental concept in modern hydrogeology and its importance is in that it allows us to assess the quality of the prediction. In the Background Study, once a decision was made to accept the normal model, it was viewed as a certain model and that does not model realistically the uncertainty.

Respectfully,

Yoram Rubin

## References

Rubin, Y., 2003, *Applied Stochastic Hydrogeology*, Oxford University Press.

Rubin, Y., X. Chen, H. Murakami, and M. Hahn, 2010, A Bayesian approach for data assimilation and conditional simulation of spatial random fields. *Water Resources Research*, 46, W10523, doi:10.1029/2009WR008799.

Shapiro, S.S., and M.B. Wilk, 1965, *Biometrika*, Vol. 52, No. 3/4, pp. 591-611.

Wilson, A., and Y. Rubin, 2002, Characterization of aquifer heterogeneity using indicator variables for solute concentrations, *Water Resour. Res.*, 38(12).