

Methods to Quantify Error Propagation and Prediction Uncertainty for USGS Raster Processing

Principal Investigators

Jason Gurdak
Hydrologist
USGS, Colorado Water Science Center
Denver Federal Center, MS 415
Lakewood, CO 80225
Phone (303) 236-4882 ext. 222
Fax (303) 236-4912
jgurdak@usgs.gov

Sharon Qi
Hydrologist/GIS Specialist
USGS, Cascades Volcano Observatory
1300 S.E. Cardinal Court, Building 10, Suite 100
Vancouver, WA 98683
Phone (360) 993-8977
Fax (360) 993-8981
slqi@usgs.gov

Primary Contact

Jason Gurdak
Hydrologist
USGS, Colorado Water Science Center
Denver Federal Center, MS 415
Lakewood, CO 80225
Phone (303) 236-4882 ext. 222
Fax (303) 236-4912
jgurdak@usgs.gov

Executive Summary

Errors associated with geospatial data can propagate through natural-science (biologic, geographic, geologic, geospatial, and hydrologic) models that utilize raster processing, resulting in significant and spatially variable prediction uncertainty. This inherent prediction uncertainty affects how model results are interpreted by scientists, environmental regulators, resource managers, elected officials, and the general public. Frequently, USGS scientists use raster processing of geospatial data to create independent variables for empirical models, boundary conditions for mechanistic models, extrapolate beyond observed data (dependent variables) to make predictions for unobserved cases or spatial extents, and to make relatively simple, everyday calculations, such as defining depth to water using land-surface and water-table raster data sets. Yet, the propagation of input errors from geospatial data and resulting prediction uncertainty of raster-based models are rarely quantified.

To maintain scientific leadership and provide the best available science, the USGS must address the following priority research questions: What role does the propagation of error from geospatial data during raster processing have on prediction uncertainty of USGS models and calculations? How can this prediction uncertainty be quantified in these USGS models? Can prediction uncertainty be minimized in future iterations of these USGS models? These priority research questions are addressed in the following proposal designed to develop and implement a stochastic-based method to identify the propagation of input errors from geospatial data during raster processing and to quantify the associated prediction uncertainty in USGS models. A novel ArcGIS tool will be developed that uses Latin Hypercube Sampling (a stratified stochastic approach similar to Monte Carlo analysis) to quantify error propagation and prediction uncertainty of geospatial models. As a demonstration example of the approach, utility, and high likelihood of success, the proposed method has been applied to a ground-water quality model of the High Plains aquifer. This application of the proposed method successfully demonstrates that spatially-variable prediction uncertainty of geospatial models can be quantified, and illustrates that errors can be evaluated to reduce this uncertainty in future iterations of the model. The demonstration example uses a common USGS hydrologic model, but the method and tool developed under this proposal would have cross-disciplinary applications for any biologic, geographic, geologic, geospatial, or hydrologic raster-based USGS model or product.

1.0 Project Description

1.1 Research Questions and Hypotheses

To maintain scientific leadership and provide the best available science to the public and cooperators, USGS scientists must seek to present our data to the best of our ability and this includes estimates of and information on uncertainty of geospatial data and associated raster-based predictive models. A major challenge is how to do this and make it understandable to USGS scientists and users of our geospatial data and models. Therefore, the following priority research questions must be addressed:

Priority Research Question 1 — What role does the propagation of input error from geospatial data during raster processing have on prediction uncertainty of raster-based USGS models and geospatial data products?

Priority Research Question 2 — How can this prediction uncertainty be quantified in these raster-based USGS models and products?

Priority Research Question 3 — How can prediction uncertainty be minimized in future iterations of these raster-based USGS models and products?

The following research hypotheses will be evaluated and tested to provide answers to the above priority research questions.

Hypothesis 1 — inherent error within geospatial data propagates through raster processing of simple and complex USGS models and results in spatially variable and potentially significant uncertainty associated with model output or predictions.

Hypothesis 2 — the stochastic modeling approach of Latin Hypercube Sampling can be successfully implemented during raster processing to quantify the propagation of input error from geospatial data and resulting prediction uncertainty.

Hypothesis 3 — the results of Latin Hypercube Sampling provide necessary information and a systematic method to reduce prediction uncertainty of future iterations of raster-based models.

1.2 Objectives

The primary objectives of the proposed research will address the previously stated research questions and hypotheses.

Objective 1 — develop an ArcGIS script and associated graphical user interface (GUI) as a tool within ArcMap that provides a stochastic (Latin Hypercube Sampling) framework to quantify the propagation of input errors and prediction uncertainty of raster processing.

Objective 2 — evaluate the utility of the tool (outlined in Objective 1) for quantifying error propagation, quantifying prediction uncertainty, and minimizing uncertainty in simple calculations to complex models, across cross-disciplinary raster processing applications.

Objective 3 — document the tool (Objective 1) in a User's Guide, and publish the findings from the Objective 2 evaluation in peer-reviewed journals.

1.3 Background and Literature Review

A fundamental activity common to all USGS disciplines is the collection of field-data (biologic, geographic, geologic, and hydrologic data) that is relevant to each science thrust. Traditionally, these data consisted of point-measurements or sample-collections at discrete locations. This traditional view of field data has been expanded with the collection and development of numerous types of geospatial data. Geospatial data represent parameters (attributes) of large geographic areas that aren't feasible to sample or measure by hand and are typically represented in a geographic information system (GIS). GIS provides a set of efficient tools for collecting, storing, retrieving, transforming, displaying, analyzing, and computationally modeling geospatial field-data (Burrough, 1986) that has wide application to and support of a majority of USGS science thrusts. Therefore, the manipulation of geospatial data within GIS plays a critically important role in the USGS Mission and Strategic Direction.

Arguably one of the most common types of manipulation of geospatial data in GIS (perhaps on a daily basis by many USGS GIS-practitioners) is the use of raster processing to make new geospatial data sets. The scope of raster processing ranges from simple to complex computational models, with the objective of creating geospatial data for either stand-alone applications or implementation in more complex natural science USGS models. The importance of raster processing is illustrated by the fact that the resulting geospatial data have a very frequent direct or indirect influence on the results of many USGS models. The popularity and far-reaching application of raster processing has created a significant challenge for USGS scientists and GIS-practitioners; endemic uncertainty associated with USGS model predictions because of inherent propagation of geospatial data error during raster processing (Mowrer and Congalton, 2000).

This challenge stems from the unavoidable and inherent errors associated with geospatial data in GIS as imperfect representations of the real world (Zhang and Goodchild, 2002; Hunsaker

et al., 2001; Burrough and McDonnell, 1998). There are two main types of GIS errors: a) source errors that exist in geospatial data; and b) error propagation through operations performed on these data, including raster processing (Yeh and Li, 2003; Heuvelink, 1998). The data source errors of geospatial data are defined by the difference between reality and the representation of the reality in the geospatial data and are a function of the accuracy and precision of the geospatial data (Mowrer and Congalton, 2000; Heuvelink, 1998; Heuvelink et al., 1989). Accuracy of geospatial data refers to the closeness of represented measurements or computations to their “true” or accepted values, and precision refers to the level of measurement and exactness of descriptions reported in the geospatial data (Gottsegen et al. 1999). Although recent research efforts have improved quantification methods for many types of source errors associated with geospatial data in GIS, there is no generally accepted theory for handling error propagation in GIS (Heuvelink, 1998). More importantly, many USGS GIS-practitioners may be aware of error propagation during raster processing, but in practice, rarely address or quantify this problem because of the lack of a universally available ArcGIS tool or methodology.

Error propagation occurs because the output of a raster process or GIS operation is a function of the input geospatial data sets, which have inherent source errors that automatically affect the computed results (Heuvelink, 1998). The cause of error propagation is generally more complex because source errors are not the only errors that propagate through raster processing. Many USGS raster processing applications use simple and complex computational models during raster processing (Gurdak and Qi, 2006; Qi and Gurdak, 2004 and 2006). The model coefficients or model structure are subject to estimation error (van Horssen et al., 2002). Therefore, the uncertainty of results from raster processing is a function of error propagation from both source errors of geospatial data and the model errors introduced by the GIS model.

Stochastic modeling, such as Monte Carlo analysis, has previously been identified as a successful method to identify error propagation and quantify uncertainty associated with other types of GIS applications (van Horssen et al., 2002; Sklar and Hunsaker, 2001; Phillips and Marks, 1996). An alternative stochastic modeling approach that will likely have widespread benefit for quantifying uncertainty associated with raster processing common to USGS applications is Latin Hypercube Sampling (Gurdak and Qi, 2006). Latin Hypercube Sampling (LHS) (McKay et al., 1979) is a widely used variation on the standard Monte Carlo (MC) stochastic sampling method for performing uncertainty analysis. The MC technique uses simple random sampling of the input probability distributions and commonly requires a large number of realizations to approximate the input probability distribution. In contrast, LHS uses a stratified

sampling technique that allows distribution of samples drawn to correspond more closely with the input probability distribution (McKay et al., 1979). For the same number of samples, the LHS correspondence produces an unbiased estimate of the mean and a smaller variance, as compared to MC. This smaller variance translates in a greater confidence, fewer model simulations, and faster computation times necessary for use within ArcMap (Gurdak and Qi, 2006). This is especially beneficial for complex USGS model simulations because running enough simulations to properly represent the input distribution may be impractical using MC.

1.4 Approach

To illustrate the utility of the proposed research, the approach is demonstrated on a recently developed ground-water vulnerability model of the High Plains aquifer (Gurdak and Qi, 2006; Qi and Gurdak, 2006). Because the proposed ArcGIS tool (Objective 1) was not available, the stochastic modeling for this demonstration was performed using the proprietary software @Risk (Palisade Corporation, 2002). This software is not coupled with GIS, making it cumbersome to integrate the error propagation analysis within raster processing. This is further motivation for the proposed Objective 1; to use a combination of Python scripting and Visual Basic (VB) to create a user friendly GUI-based (Graphical User Interface) tool for implementing Latin Hypercube Sampling to quantify error propagation during raster processing. Although the following demonstration example of the proposed approach is specific to a hydrologic model and a complex raster processing example, the proposed ArcGIS tool will enable the following analysis for more general and cross-disciplinary raster processing applications.

1.4.1 Demonstration Example

Gurdak and Qi (2006) developed a vulnerability model that predicts the probability of detecting nitrate concentrations greater than 4 mg/L in ground water of the High Plains aquifer. This model was expressed as a ground-water vulnerability map (Figure 1) created using GIS raster processing (map algebra). This research demonstrated that the propagation of input error was significant and resulted in spatially variable prediction uncertainty in this ground-water vulnerability map. The approach used to create the resulting prediction uncertainty map (Figure 2) during GIS raster processing is outlined below.

Latin Hypercube Sampling (LHS) was used to develop the uncertainty prediction intervals, which defines the error range surrounding the estimates of predicted probability of ground-water vulnerability to nitrate concentrations greater than 4 mg/L in ground-water within each 80-m GIS grid cell. The 90% uncertainty prediction interval range was reported (Figure 2),

and is defined by the difference between the 5th and 95th percentile of the output probability distribution of the LHS modeling, and represents the likelihood that the true predicted probability of ground-water vulnerability to nitrate greater than 4 mg/L is within that uncertainty prediction interval. Because the input errors, and thus the propagated model output uncertainty, typically are spatially variable (Phillips and Marks 1996), the uncertainty was calculated at each GIS grid cell during raster processing and the 95% uncertainty prediction intervals were presented as uncertainty maps (Figure 2) to accompany the final vulnerability map (Figure 1). For each model simulation, 1,000 Latin hypercube sampling iterations were run. As suggested by Phillips and Marks (1996), all input probability distributions were assumed normal; each distribution mean was assigned as the estimated model coefficient or attributed value for that GIS grid cell. The estimation variance of the input probability distributions was defined by a conservative range of source errors for each geospatial data and by the Wald 95% confidence intervals for ground-water vulnerability model coefficients during the raster processing.

Source errors are typically not available for many geospatial data sets; however, reasonable estimates of errors were obtained for the geospatial data used in this demonstration example. As a first approximation, the source errors were estimated to range from 10 to 28% and were obtained from various sources. For example, explanatory variable error equal to 20% was used for proportion of irrigated agricultural land, based on Qi et al. (2002) use of satellite imagery from Landsat Thematic Mapper (nominal date 1992) and raw National Land Cover Data (NLCD) satellite data to classify irrigated and non-irrigated land. Qi et al. (2002) used ground-reference information from 2,500 km² for comparison against the classified irrigated land data and reported an approximate 80% correct classification and 20% error estimate. Gurdak and Qi (2006) created unsaturated-zone lithology GIS datasets by interpolating 56,000 lithologic logs from wells across the High Plains using ordinary kriging. Source error for this lithology data was estimated at 28% from the average root-mean-squared prediction error of cross validation during ordinary kriging.

A modified method originally presented by van Horssen et al. (2002) to evaluate spatial interpolation during ordinary block kriging is proposed as a method to evaluate uncertainty contributions from various sources during raster processing, by comparing the relative variance contributions. For the outcome of raster processing at any given GIS grid cell, represented as X , the total prediction variance of the raster process, $\sigma^2(X)$, is equal to the sum of the variance as a result of uncertainty in the raster model errors, $\sigma_m^2(X)$, and variance as a result of uncertainty in source errors of geospatial data, $\sigma_d^2(X)$. The decomposition of the total prediction variance is:

$$\sigma^2(X) = \sigma_m^2(X) + \sigma_d^2(X) + 2 \times Cov\left[\sigma_m^2(X), \sigma_d^2(X)\right] \quad (1)$$

where $2 \times Cov\left[\sigma_m^2(X), \sigma_d^2(X)\right]$ is twice the covariance of each pair of terms formed from the components of the sum (Hosmer and Lemeshow 2000). Because the calculated covariance of each pair of terms is typically negligible, it is reasonable to assume independence between the other variance components (Gurdak and Qi, 2006). Therefore, the relative variance contribution due to model errors (RVC_m) is calculated as:

$$RVC_m = \frac{\sigma_m^2(X)}{\sigma^2(X)} \times 100 \% \quad (2)$$

and the relative variance contribution due to geospatial data source errors (RVC_d) is:

$$RVC_d = \frac{\sigma_d^2(X)}{\sigma^2(X)} \times 100 \% \quad (3)$$

If uncertainty due to model error and source errors of geospatial data contributes equally, the RVC_m will equal the RVC_d. RVC_m values greater than the RVC_d values indicate locations of the study area where raster processing prediction uncertainty due to model errors dominates the total uncertainty. This result would indicate the need for improved model development, such as additional monitoring wells to better characterize the variability of nitrate concentration in this demonstration example. RVC_m values lower than the RVC_d values indicate locations of the study area where uncertainty due to source errors in geospatial data dominates the total uncertainty. Therefore, these locations represent where improved measurement accuracy and precision of GIS-based explanatory variables are needed to reduce RVC_d and thus improve raster prediction uncertainty.

Results of this demonstration example illustrate that prediction uncertainty from raster processing can be significant and spatially variable (Figure 2). Prediction uncertainty of the ground-water vulnerability model results ranges from 0 to 100%, with generally greater uncertainty in the southern and central study area as compared to the northern study area (Figure 2). Comparison of the RVC values reveals that the model errors constitute the majority of the overall raster prediction uncertainty (Gurdak and Qi, 2006). Spatial patterns of uncertainty contributions emerge; prediction uncertainties across the southern and central study area are largely due to model errors (i.e., a lack of monitoring wells needed to most adequately describe the spatial variability of nitrate concentrations). A systematic and cost-effective strategy to

reduce prediction uncertainty in the demonstration model may be to add monitoring wells to the southern and central areas, specifically in areas with the widest prediction intervals, identified in Figure 2. Conversely, the RVC_d of the northern study area is significantly larger than RVC_m , indicating that source errors of geospatial data are greater than errors caused by model errors. Therefore, a cost-effective strategy to reduce prediction uncertainty of the northern area would entail acquiring more accurate input geospatial data sets (Figure 2).

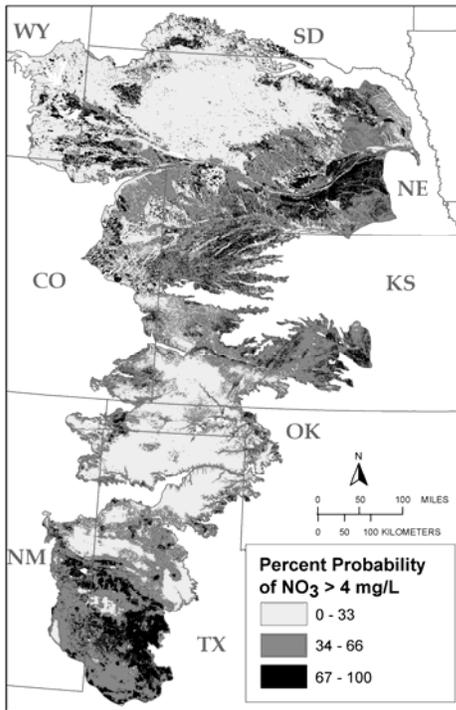


Figure 1. A ground-water vulnerability map; illustrating the spatial distribution of the raster-based predicted probability of detecting nitrate >4 mg/L in ground water of the High Plains aquifer.

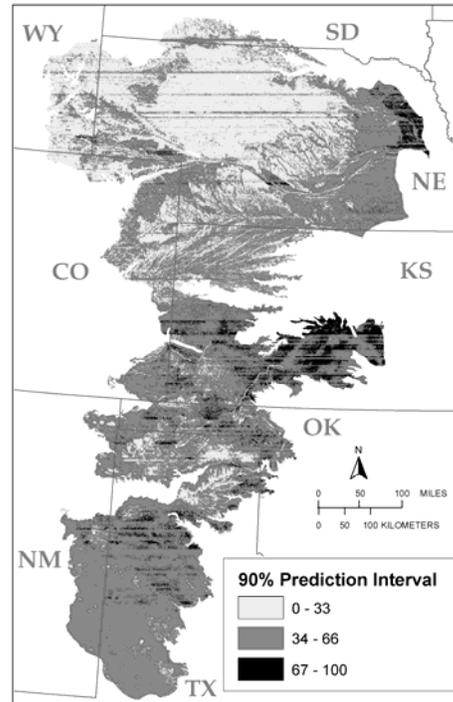


Figure 2. An uncertainty map; illustrating the spatial uncertainty of predictions from the ground-water vulnerability model and map (Figure 1), which was developed using the proposed approach.

1.5 Expected Results and Products

Several specific products will be generated from this research. The primary and most relevant product to the USGS mission will be the GUI-based tool for ArcMap that provides a Latin Hypercube Sampling framework to quantify the propagation of input errors and prediction uncertainty during raster processing. This tool will have widespread and cross-disciplinary application. A conceptual graphic of what this would look like is illustrated in Figure 3. This tool will also provide an option for the Relative Variance Contribution (RVC) calculations, which can be used to reduce source error and model error and improve overall confidence in results

from raster processing. A User's Guide for this tool will be published in a USGS report series. In addition, at least two publications are anticipated for publication in peer-reviewed journals. Upon completion of this research, an internal project report summarizing major findings and results will be written and submitted to the CEGIS.

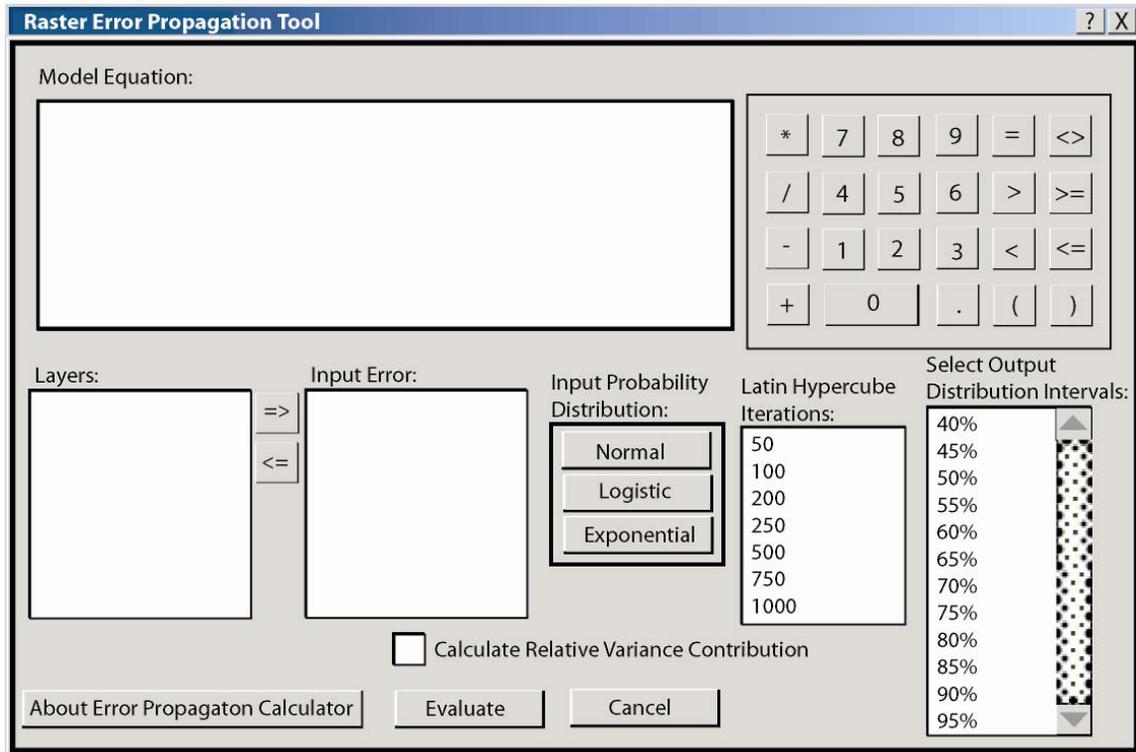


Figure 3. A conceptual illustration of the GUI for the proposed ArcGIS tool (tentatively named “Raster Error Propagation Tool”) to identify error propagation and quantify prediction uncertainty associated with raster processing of USGS models.

1.6 Significance to the USGS Mission

Inherent errors associated with geospatial data propagate through all USGS models and calculations that utilize raster processing in GIS, which can result in significant and spatially variable prediction uncertainty. The resulting uncertainties of USGS model predictions bear societal consequences and have significant implications for the USGS mission. Prediction uncertainty affects how model results are interpreted by scientists, environmental regulators, resource managers, elected officials, and the general public. Thus, the most direct benefit of this research to the USGS mission is the proposed tool (outlined in Objective 1 and Figure 3), which will provide USGS GIS-practitioners a flexible methodology to quantify error propagation, quantify prediction uncertainty, and a method to reduce prediction uncertainty in future iterations of raster-based USGS models. Furthermore, the development and availability of such a tool will

increase the awareness of USGS GIS-practitioners of the potential limitations to raster processing. Ideally, spatial metadata should provide GIS users with information about a variety of errors that are inherent in the raster data sets. However, such information is rarely available for geospatial data developed and used by USGS GIS practitioners. Therefore, this proposed tool will not only compliment current USGS research activities to better quantify and report error in spatial metadata, but will also serve as motivation for all USGS scientists that use GIS raster processing to address and quantify the inherent source errors of geospatial data.

2.0 References

- Burrough, P.A., and McDonnell, R.A., 1998. Principles of geographical information systems. Oxford University Press, New York, 333 p.
- Burrough, P.A., 1986. Principles of geographical information systems for land resources assessment, Oxford, Clarendon Press.
- Gar-On Yeh, A., and Li, X., 2003. Error propagation and model uncertainties of cellular automata in urban simulation with GIS. GeoComputation, conference proceedings of the 7th International conference on GeoComputation, United Kingdom, www.geocomputation.org/2003/Papers/Yeh_And_Li_Paper.pdf
- Gottsegen, J., D. Montello, and M. Goodchild. 1999. A comprehensive model of uncertainty in spatial data. In Lowell, K., and Jaton, A., eds., Spatial accuracy assessment: Land information uncertainty in natural resources. Chelsea, Michigan: Ann Arbor Press.
- Gurdak, J.J., and S.L. Qi. 2006. Vulnerability of recently recharged ground water in the High Plains aquifer to nitrate contamination. USGS SIR 2006-5050.
- Heuvelink, G.B.M., 1998. Error propagation in environmental modeling with GIS. London: Taylor & Francis Ltd.
- Heuvelink, G.B.M., and Burrough, P.A., 1993. Error propagation in cartographic modeling using Boolean logic and continuous classification. International Journal of Geographical Information Systems, 7(3), 213-246.
- Heuvelink, G.B.M., Burrough, P.A., and Stein, A., 1989. Propagation of errors in spatial modeling with GIS. Intern. J. of Geographical Information Systems, 3(4), 303-322.
- Hosmer, D.W., and S. Lemeshow. 2000. Applied logistic regression. N.Y., John Wiley & Sons.
- Hunsaker, C.T., Goodchild, M.F., Friedl, M.A., and Case, T.J., 2001. Spatial uncertainty in ecology, Implications for remote sensing and GIS applications. Springer, N.Y. 402 p.
- McKay, M.D., R.J. Beckman, and W.J. Conover. 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21: 239–245.
- Mowrer, H.T., and Congalton, R.G., 2000. Quantifying spatial uncertainty in natural resources: Theory and applications for GIS and remote sensing. Ann Arbor Press, 244 p.
- Palisade Corporation. 2002. Guide to using @RISK, risk analysis and simulation add-in for Microsoft Excel. Newfield, New York: Palisade Corporation.
- Phillips, D.L., and D.G. Marks. 1996. Spatial uncertainty analysis: propagation of interpolation errors in spatially distributed models. Ecol. Modeling, 213-229.
- Qi, S.L., A. Konduris, D.W. Litke, and J. Dupree. 2002. Classification of irrigated land using satellite imagery, the High Plains aquifer, nominal data 1992. USGS Water-Resources Investigations Report 02–4236.
- Qi, S.L., and J.J. Gurdak. 2004. GIS and statistical groundwater vulnerability modeling. ESRI International User Conference Proceedings 2004.

- Qi, S.L., and J.J. Gurdak. 2006. Percentage of probability of nonpoint source nitrate contamination of recently recharged ground water in the High Plains aquifer, USGS Data Series, available at http://water.usgs.gov/lookup/getspatial?ds192_hp_npctprob.
- Sklar, F. H., and Hunsaker, C.T., 2001. The use and uncertainties of spatial data for landscape models: An overview with examples from the Florida Everglades. In Hunsaker, C.T., Goodchild, M.F., Friedl, M.A., and Case, T.J., eds., *Spatial uncertainty in ecology, Implications for remote sensing and GIS applications*. Springer. New York. 402 p.
- van Horssen, P.W., E.J. Pebesma, and P.P. Schot. 2002. Uncertainties in spatially aggregated predictions from a logistic regression model. *Ecological Modelling* 154, no. 1-2: 93–101.
- Zhang, J., and Goodchild, M.F., 2002. *Uncertainty in geographical information*. Taylor and Francis, London, 266 p.
- Zhang, R., J.D. Hamerlink, S.P. Gloss, and L. Munn. 1996. Determination of non-point source pollution using GIS and numerical models. *J. of Environmental Quality* 25,411–418.

3.0 Project Support

The research outlined in this proposal has a high likelihood of success because of a strong collaboration and combined expertise between researchers at the USGS Colorado Water Science Center, Lakewood, CO and at the USGS Cascades Volcano Observatory, Vancouver, WA. Although this proposal has identified no additional (internal or external) support, the high likelihood of success of this research is leveraged by prior internal USGS support from the National Water Quality Assessment (NAWQA) Program. NAWQA support during this past fiscal year enabled the research that resulted in the necessary foundation and background of the proposed research. Future NAWQA funding is not available to support the proposed research.

4.0 Budget Request

Fiscal Year 2007 Budget	Budget amount for cost center 8582	Total Year 1
Personnel Salary (Gurdak and Qi)	81,000	81,000
Publication Costs	4,000	4,000
TOTAL DIRECT	85,000	85,000
Gross Assessment Rate	58.2%	58.2%
INDIRECT COSTS ESTIMATE	49,500	49,500
TOTAL	\$134,500	\$134,500