Abstract—This paper evaluates heterogeneous information fusion using multi-task Gaussian processes in the context of geological resource modeling. Specifically, it empirically demonstrates that information integration across heterogeneous information sources leads to superior estimates of all the quantities being modeled, compared to modeling them individually. Multi-task Gaussian processes provide a powerful approach for simultaneous modeling of multiple quantities of interest while taking correlations between these quantities into consideration. Experiments are performed on large scale real sensor data.

I. INTRODUCTION

In applications such as space-exploration, mining or agriculture automation, modeling the underlying resource is a fundamental problem. For such applications, an efficient, flexible and high-fidelity representation of the geology is critical. The key challenges in realizing this are that of dealing with the problems of uncertainty and incompleteness. Uncertainty and incompleteness are virtually ubiquitous in any sensor based application as sensor capabilities are limited. The problem is magnified in a field automation scenario due to sheer scale of the application. Incompleteness is a major problem in any large scale resource modeling endeavor as sensors have limited range and applicability. A more significant contributor to this issue is that of cost - sampling and collecting such data is expensive. Geological data is typically collected through various sensors/processes of widely differing characteristics and consequently lead to different kinds of information. Often the resource is characterized by numerous quantities (for example, soil composition in terms of numerous elements). These quantities often are correlated.

Given these issues, large scale geological resource modeling needs a representation that can handle spatially correlated, incomplete and uncertain data. Not only must the correlation between homogeneous quantities be modeled but also that between heterogeneous quantities. This paper uses a Gaussian process (GP) representation of resource data similar to that described in [1]. GPs are ideally suited to handling spatially correlated data. This paper further uses an extension of the basic Gaussian process model, the multi-task Gaussian process (MTGP) which is used to simultaneously model multiple quantities of interest. The proposed model not only captures spatial correlations between individual quantities with themselves (at different locations) but also that between totally different quantities that together quantify the resource. That the quantities modeled in this paper exhibit strong correlation is known from geological sciences. This paper presents an empirical evaluation to understand (1) if simultaneous modeling of multiple quantities of interest (i.e. modeling and using the correlations between them and hence performing data fusion) is better than modeling these quantities independently and (2) if the nonstationary kernels are more effective than stationary kernels at modeling geological data. Experiments are performed on large scale real sensor data.

II. RELATED WORK

Gaussian processes [2] (GPs) are powerful non-parametric Bayesian learning techniques that can handle correlated, uncertain and incomplete data. They have been used in a range of fields [3], [1]. They produce a scalable multi-resolution model of the entity under consideration. They yield a continuous domain representation of the data and hence can be sampled at any desired resolution. They incorporate and handle uncertainty in a statistically sound manner and represent spatially correlated data appropriately. They model and use the spatial correlation of the given data to estimate the values for other unknown points of interest. In an estimation sense, GPs provide the best linear unbiased estimate [4] based on the underlying stochastic model of the spatial correlation between the data points. They basically perform an interpolation methodology called Kriging [5] which is a standard interpolation technique.

The work [1], performed in the context of terrain modeling, also proposed the use of non-stationary kernels (neural network) to model large scale discontinuous spatial data. It compared performances of GPs based on stationary (squared exponential) and non-stationary (neural network) kernels as well as several other standard interpolation methods applicable to alternative representations of terrain data. The nonstationary neural network kernel was found to be superior to the stationary squared exponential kernel and at least as good as most standard interpolation techniques for a range of terrain (in terms of sparsity/complexity/discontinuities). The work presented in this paper builds on this GP representation. However, it addresses the problem of simultaneous modeling multiple heterogeneous quantities of interest. This requires the modeling and usage of the correlations between these quantities towards improving predictions of each of them - an instance of data fusion using Gaussian processes.

Two preliminary attempts towards data fusion using Gaussian processes include [6] and [7]. The former bears a “hierarchical learning” flavor to it in that it demonstrates how a GP can be used to model an expensive process by (a) modeling a GP on an approximate or cheap process and (b)
using the many input-output data from the approximate process and the few samples available of the expensive process together in order to learn a GP for the latter. The latter work attempts to generalize arbitrary transformations on GP priors through linear transformations. It hints at how this framework could be used to introduce heteroscedasticity (random variables with non-constant variance) and how information from different sources could be fused. However, specifics on how the fusion can actually be performed are beyond the scope of the work. The work [8] integrates heterogeneous information within a Gaussian process classification setting, in a protein fold recognition application domain. Each feature representation is represented by a separate GP. The fusion uses the fact that individual feature representations are considered independent and hence a composite covariance function would be defined in terms of a linear sum of Gaussian process priors. This paper addresses the data fusion problem wherein multiple, possibly heterogeneous sources of information are correlated. The objective is to use this correlation to improve the estimate of the quantity being modeled (geological concentrations in this paper). A recent work [9] integrates “hard” data obtained from sensors with “soft” information from human sources within a Gaussian process classification framework, by using a separate kernel for each data type and combining all the kernels using a product rule. This problem/approach is different from the work presented here. It uses heterogeneous information sources as mutually independent sources of information that are transformed into the kernel representation and combined using a product rule. The approach presented in this paper improves the estimate of the quantity (or quantities) being modeled by explicitly modeling the correlation between multiple heterogeneous information sources.

Two recent approaches demonstrating data fusion with Gaussian processes in the context of large scale terrain modeling were based on heteroscedastic GPs (HGPS) [10] and dependent GPs (DGPs) ([11] and [12]). These address the problem of fusing multiple, multi-sensor data sets of a single quantity of interest. This paper describes the framework for extending this concept to multiple heterogeneous quantities of interest. It performs a empirical study to quantify the benefits (if any) of simultaneous modeling of the multiple quantities by modeling and using the correlations between them as against modeling each of these quantities separately. This paper also compares data fusion with both the non-stationary neural network and stationary squared exponential kernels in the context of modeling geological data.

III. APPROACH

A. Gaussian processes

Gaussian processes ([2]) (GPs) are stochastic processes wherein any finite subset of random variables is jointly Gaussian distributed. They may be thought of as a Gaussian probability distribution in function space. They are characterized by a mean function \(m(x)\) and the covariance function \(k(x,x')\) that together specify a distribution over functions. In the context of geological resource modeling, each \(x \equiv (\text{east}, \text{north}, \text{depth})\) (3D coordinates) and \(f(x) \equiv z\), the concentration of the quantity being modeled. Although not necessary, the mean function \(m(x)\) may be assumed to be zero by scaling/shifting the data appropriately such that it has an empirical mean of zero.

The covariance function or kernel models the relationship between the random variables corresponding to the given data. It can take numerous forms (see chapter 4 in [2]). The stationary squared exponential (or Gaussian) kernel (SQEXP) is given by

\[ k_{\text{SQEXP}}(x, x', \Sigma) = \sigma_f^2 \exp \left( -\frac{1}{2} (x - x')^T \Sigma (x - x') \right) \]

(1)

where \(k\) is the covariance function or kernel; \(\Sigma = \text{diag}[\sigma_{\text{east}}, \sigma_{\text{north}}, \sigma_{\text{depth}}]^{-2}\) is a \(d \times d\) diagonal length-scale matrix \((d = \text{dimensionality of input } = 3 \text{ in this case})\), a measure of how quickly the modeled function changes in the east, north and depth directions; \(\sigma_f^2\) is the signal variance. The set of parameters \(\{\sigma_{\text{east}}, \sigma_{\text{north}}, \sigma_{\text{depth}}, \sigma_f\}\) are referred to as the kernel hyperparameters. The non-stationary neural network (NN) kernel ([13], [14] and [15]) takes the form

\[ k_{\text{NN}}(x, x', \Sigma) = \sigma_f^2 \frac{2}{\pi} \arcsin \left( \frac{2\sqrt{\Sigma} x x'}{(1 + 2\sqrt{\Sigma} x x')^{\frac{3}{2}}} \right) \]

(2)

where \(\tilde{x}\) and \(\tilde{x}'\) are augmented input vectors (each point is augmented with a 1), \(\Sigma = \text{diag}[\beta, \sigma_{\text{east}}, \sigma_{\text{north}}, \sigma_{\text{depth}}]^{-2}\) is the \((d + 1) \times (d + 1)\) diagonal length-scale matrix (as before) with \(\beta\) being a bias factor and \(d\) being the dimensionality of the input data. The variables \(\{\beta, \sigma_{\text{east}}, \sigma_{\text{north}}, \sigma_{\text{depth}}, \sigma_f\}\) constitute the kernel hyperparameters. The NN kernel represents the covariance function of a neural network with a single hidden layer between the input and output, infinitely many hidden nodes and using a Sigmoidal transfer function [14] for the hidden nodes. Hornik in [16] showed that such neural networks are universal approximators and Neal [13] observed that the functions produced by such a network would tend to a Gaussian process. Prior work ([11]) has found the NN kernel to be more effective than the SQEXP kernel at modeling discontinuous data.

Regression using GPs uses the fact that any finite set of training (evaluation) data and test data of a GP are jointly Gaussian distributed. Assuming noise free data, this idea is shown in Expression 3 (hereafter referred to as Equation 3). This leads to the standard GP regression equations yielding an estimate (the mean value, given by Equation 4) and its uncertainty (Equation 5).

\[
\begin{bmatrix}
    z \\
    f_s
\end{bmatrix}
\sim N \left( \begin{bmatrix}
    0 \\
    0
\end{bmatrix}, \begin{bmatrix}
    K(\cdot, \cdot) & K(\cdot, \cdot) \\
    K(\cdot, \cdot) & K(\cdot, \cdot)
\end{bmatrix} \right)
\]

(3)

\[
\begin{bmatrix}
    z \\
    f_s
\end{bmatrix}
= K(\cdot, \cdot) K(\cdot, \cdot)^{-1} z
\]

(4)

\[
\text{cov}(f_s) = K(\cdot, \cdot) - K(\cdot, \cdot) K(\cdot, \cdot)^{-1} K(\cdot, \cdot)
\]

(5)
For \( n \) training points \((X, z) = (x_i, z_i)_{i=1...n}\) and \( n^* \) test points \((X^*, f^*)\), \( K(X, X^*) \) denotes the \( n \times n^* \) matrix of covariances evaluated at all pairs of training and test points. The terms \( K(X, X), K(X^*, X^*) \) and \( K(X, X^*) \) are defined likewise. In the event that the data being modeled is noisy, a noise hyperparameter (\( \sigma \)) is also learnt with the other GP hyperparameters and the covariance matrix of the training data \( K(X, X) \) is replaced by \( [K(X, X) + \sigma^2 I] \) in Equations 3, 4 and 5. The hyperparameters are learnt by maximizing the marginal likelihood of the observed training data ([2], [1]), described in Equation 6

\[
\log p(z|X, \theta) = -\frac{1}{2}z^T K(X, X)^{-1}z - \frac{1}{2} \log |K(X, X)| - \frac{n}{2} \log(2\pi) \tag{6}
\]

B. Multi-task Gaussian processes (MTGPs)

The problem being addressed in this paper can be described as follows. The objective is to model multiple heterogeneous quantities (e.g. concentrations of various elements) of the entity in consideration (e.g. land mass). The data fusion aspect of this problem is the improved estimation of each one of these quantities by integration or use of all other quantities of interest. If each quantity was modeled using a separate GP, the objective is to improve one GP’s prediction estimates given all other GP models.

Let the number of outputs/tasks that need to be simultaneously modeled be denoted by \( nt \). Equations 3, 4 and 5 represent respectively the DGP data fusion model, the regression estimates and their uncertainties, subject to the following modifications to the basic notation. The set \( z = [z_1, z_2, z_3, \ldots, z_{nt}]^T \) represents the output values of the selected training data from the individual \( nt \) tasks that need to be simultaneously modeled. The term \( X = [X_1, X_2, X_3, \ldots, X_{nt}] \) denotes the input location values (east, north, depth) of the selected training data from the individual data sets. Any kernel ([2]) may be used and even different kernel could be used for different data sets using the technique demonstrated in [17] (for stationary kernel) or the convolution process technique demonstrated in [18], [19], [11], [12] and in this paper (for both stationary and nonstationary kernel). The covariance matrix of the training data is given by

\[
K(X, X) = \begin{bmatrix}
K_{11}^Y & K_{12}^Y & \cdots & K_{1nt}^Y \\
K_{21}^Y & \ddots & \cdots & \vdots \\
\vdots & \ddots & \ddots & \vdots \\
K_{nt1}^Y & \cdots & \cdots & K_{ntnt}^Y
\end{bmatrix}
\]

where

\[
K_{ii}^Y = K_{ii}^{U}(x_i, x_i) + \sigma_i^2 I \\
K_{ij}^Y = K_{ij}^{U}(x_i, x_j)
\]

Here, \( K_{ii}^Y \) represents the auto-covariance of the \( i^{th} \) data set with itself and \( K_{ij}^Y \) represents the cross covariance between the \( i^{th} \) and \( j^{th} \) data sets. These terms model the covariance between the noisy observed data points (\( z \) values). Thus, they also take the noise components of the individual data sets / GPs into consideration. The corresponding noise free terms are respectively given by \( K_{ii}^{U} \) and \( K_{ij}^{U} \). These are derived by using the process convolution approach to formulating Gaussian processes; details of this follow in the subsequent paragraphs. The covariance matrix between the test points and training points is given by

\[
K(X, X) = [K_{i1}^{U}(x_i, x_1), K_{i2}^{U}(x_i, x_2), \ldots, K_{i\text{nt}}^{U}(x_i, x_{nt})], \\
\text{where } i \in \{1 \ldots nt\} \text{ is the GP that is being evaluated given all other GPs. The matrix } K(X, X) \text{ is defined likewise. Finally, the covariance of the test points is given by}
\]

\[
K(X, X^*) = K_{ii}^{U}(x_i, x_i*) + \sigma_i^2 I,
\]

assuming the \( i^{th} \) GP needs to be evaluated for the particular test point. The mean and variance of the concentration estimate can thus be obtained by applying Equations 4 and 5, after incorporating multiple outputs/tasks, multiple GP/noise hyperparameters and deriving appropriate auto and cross covariances functions that model the spatial correlation between the individual data sets. Data fusion is thus achieved.

![Fig. 1. A simple demonstration of the DGP / MTGP concept. Two sine waves (black) are to be modeled. One is an inverted version of the other. Further noisy samples are available all over one of them (red) whereas the other one has noisy samples only in one part of it (blue). Merely using these few samples would result in a poor prediction of the sine wave in the areas devoid of samples. Using the spatial correlation with the red sampled sine wave enables the DGP/MTGP approach to improve the prediction of the blue sampled sine wave.](image-url)
in the MTGP approach by correlating individual heterogeneous outputs/tasks and using this correlation information to improve the prediction estimates of each of them.

The process convolution approach ([18]) is a generic methodology which formulates a GP as a white noise source convolved with a smoothing kernel. Modeling the GP then amounts to modeling the hyperparameters of the smoothing kernel. The advantage of formulating GPs this way is that it readily allows the GP to be extended to model more complex scenarios, one such scenario being the multi-task or dependent GPs (MTGPs or DGPs). The work [11] used DGPs with the SQEXP kernel and was based on [19]. This work derived the auto and cross covariances for the SQEXP kernel, through a convolution integral, as a kernel correlation. Taking inspiration from [18] and [19], the work [12] derived the auto and cross covariance functions for the non-stationary NN kernel. The expressions for the auto and cross covariance functions for both kernel are provided below. Detailed explanations and derivations on the approach and its application to either kernel may be obtained from the aforementioned references.

Assume two GPs \( N(0, k_i) \) and \( N(0, k_j) \), based on either the SQEXP or the NN kernel, with with length scale matrices \( \Sigma_i \) and \( \Sigma_j \). For the stationary SQEXP kernel, the cross and auto covariances are given by Equations 7 and 8 respectively. The corresponding expressions for the nonstationary NN kernel are given by Equations 9 and 10 respectively.

\[
K^U_{ij}(x, x') = K_f(i, j) \cdot \left(2\pi\right)^{\frac{d}{2}} |\Sigma_i + \Sigma_j|^{-\frac{1}{2}} \cdot \exp\left(-\frac{1}{2}(x - x')^T \Sigma_{ij}^{-1} (x - x')\right)
\]

(7)

where

\[
\Sigma_{ij} = \Sigma_i(\Sigma_i + \Sigma_j)^{-1} \Sigma_j = \Sigma_j(\Sigma_i + \Sigma_j)^{-1} \Sigma_i
\]

\[
K^U_{ii}(x, x') = K_f(i, i) \cdot \left(\pi\right)^{\frac{d}{2}} |\Sigma_i|^{-\frac{1}{2}} \cdot \exp\left(-\frac{1}{4}(x - x')^T \Sigma_i^{-1} (x - x')\right)
\]

(8)

\[
K^U_{ij}(x, x') = K_f(i, j) \cdot 2^{\frac{d+1}{2}} |\Sigma_i|^{\frac{1}{2}} |\Sigma_i + \Sigma_j|^{-\frac{1}{2}} |\Sigma_j|^{\frac{1}{2}} \cdot k_{NN}(x, x', \Sigma_{ij})
\]

(9)

where

\[
\Sigma_{ij} = 2 \Sigma_i(\Sigma_i + \Sigma_j)^{-1} \Sigma_j
\]

\[
K^U_{ii}(x, x') = K_f(i, i) \cdot k_{NN}(x, x', \Sigma_i)
\]

(10)

In Equations 9 and 10, the term, \( k_{NN}(x, x', \Sigma_{ij}) \), is the NN kernel for two data \( x, x' \) and length scale matrix \( \Sigma_{ij} \). It is given by Equation 2, excluding the signal variance term (\( \sigma_f^2 \)). The \( K_f \) terms in Equations 7, 8, 9 and 10 are inspired by [20]. This term models the task similarity between individual tasks. Incorporating it in the auto and cross covariances provides additional flexibility to the multi-task GP modeling process. It is a symmetric matrix of size \( nt \times nt \) and is learnt along with the other GP hyperparameters. Thus, the hyperparameters of the system that need to be learnt include \( (nt \times nt + 1)/2 \) task similarity values, \( nt \cdot 2 \) or \( nt \cdot 3 \) length scale values respectively for the individual SQEXP or NN kernels and \( nt \) noise values corresponding to the noise in the observed data sets. Learning these hyperparameters by adapting the GP learning procedure described before (see Equation 6 for multiple outputs/tasks ([11], [12])).

IV. Experiments

Fig. 2. The geological resource data set. Figures 2(a), 2(b) and 2(c) respectively show the concentrations of three elements over the region of interest. The central region of points is surrounded by sparse sets of points which are not pre-filtered when applying the proposed algorithm.

Experiments were conducted on a large scale geological resource data set made up of real sensor data. The data consists of 63,667 measurements from a 3478.4 m x 1764.6 m x 345.9 m region in Australia that has undergone drilling and chemical assays to determine its composition. The holes are generally 25-100m apart and tens to hundreds of meters deep. Within each hole, data is collected at an interval of 2m. The measurements include the (east, north, depth) position data along with the concentrations of three elements, Element-1, Element-2 and Element-3, hereafter denoted as E1, E2 and E3 respectively. These three quantities are known to be correlated and hence the objective is to use each of their GP models to improve the other’s prediction estimates by capturing the correlation between these quantities. The data set is shown in Figure 2. The methodology of testing is
described in Section IV-A. Multiple metrics have been used to evaluate the methods, these are described in Section IV-B. Results obtained are then presented and discussed in Section IV-C. Outputs of the data fusion process provided by the best performing model as suggested by the evaluation are also presented.

A. Testing procedure

The objective of the experiment was to compare the multi-task GP approach with a conventional GP approach and quantify if the data fusion in the MTGP actually improves estimation. A second objective of the experiments was to compare the nonstationary NN kernel with the stationary SQ-EXP kernel. Towards these goals, a ten fold cross validation experiment was performed on the data set, with each of the two kernels. This was motivated by the work [21], which suggests a ten fold stratified (similar number of samples in each fold) cross validation as the best way of testing the estimation accuracy of machine learning methods on real world data sets.

The MTGP and simple GP approaches each require an optimization step for model learning. The optimization step in each method can result in different local minima in each trial (and with each kernel). Thus, to do a one-on-one comparison between the two approaches and quantify their relative performances, an exact comparison is required. The benchmarking experiment presented in this paper provides an exact comparison between the MTGP and GP approaches. To do this,

- The best available MTGP parameters were found for each kernel. From this, appropriate subsets of the parameters were chosen for the GP approach.
- The approaches were compared on identical test points and identical training/evaluation points selected for each of the test points.
- It is also necessary that the covariance function for the simple GP approach must be identical to the auto-covariance function of the DGP approach. For this reason, the auto-covariance function (for both kernels) is used as the covariance function for the GP approach to data fusion.

In addition to this, three independent GPs (denoted as GPI hereafter) were optimized for E1, E2 and E3 and their estimates for the same set of test points were also compared. Thus the effect of information integration in the context of the geological resource modeling can be seen in terms of both an exact comparison (MTGP vs GP) and an independent comparison (MTGP vs GPI).

For the cross validation, a “block” sampling technique was used, a 3D version of the “patch” sampling method used in [1] (see Figure 3). The idea was that rather than selecting test points uniformly, blocks of data test the robustness of the approach better as the support points to the query point are situated farther away than in uniform point selection. The data set is gridded into blocks of different sizes. Collections of blocks represent individual folds. In each cross validation test, one fold was designated as a test fold and points from it were used exclusively for testing. All other folds together constituted the evaluation data, a small subset of which were labeled as the training data. Note that this technique of testing will naturally lead to larger errors. For the test fold, the E1, E2 and E3 concentrations (and error metrics defined in the following section) are estimated first using the MTGP approach, then with the GP approach using parameters from the optimized MTGP parameters and finally, with an independently optimized GP for each of the three quantities. The result of a 10 fold cross validation test is a 63,667 point evaluation in tougher test conditions than what would be obtainable with uniform sampling (e.g. every tenth point) of test points.

Block sizes were chosen empirically, in proportion to the dimensions of the whole data set and with a view of performing a stratified cross validation test. The block sizes chosen and the resulting implications on the cross validation testing are shown in Table I. The smaller block size of 87m x 45m x 9m results in each fold having a similar number of points (i.e. numbers of points in folds with min/max test points are similar) and thus results in the most stratified cross validation test. With increasing block size, prediction error increases (support data is farther away), stratification is reduced and hence, variance in prediction error also increases. Uniform sampling of test points may be considered as a limiting case of block sampling with the smallest block size possible.

B. Metrics

Multiple metrics have been used to understand the various methods being tested. They are briefly described below. These are evaluated for each test point in each fold of the cross validation test. The result would then be represented by the mean and standard deviations of all values across all folds.

1) Squared Error (SE): This represents the squared difference between the predicted concentrations and the known concentrations for the set of test points. The mean over the set of all test points (Mean Squared Error or MSE) is the most popular metric for the context of this paper. Referring Equations 4 and 5, for
the \( i^{th} \) test point,
\[
SE(i) = (\bar{f}(i) - z_i)^2
\]

2) **Variance (VAR):** This represents the variance (uncertainty) in the predicted concentrations for the set of test points. A lower VAR is a good outcome, only if the SE is also low. A model that has high SE and low VAR would be a poor model as this result would suggest that the model is confident of its inaccurate estimates. A better outcome would be a model with high SE and correspondingly high VAR i.e. a model that has inaccurate predictions but is also uncertain about these predictions.

3) **Negative log probability / Log loss (NLP):** Inspired by [2] (see page 23), this is a measure of the extent to which the model (including the GP model, kernel, parameters and evaluation data) explain the current test point. The lower the value of this metric, the better the model. For the \( i^{th} \) test point,
\[
NLP(i) = \frac{1}{2} \log(2\pi\sigma^2) + \frac{(\bar{f}(i) - z_i)^2}{2\sigma^2(i)^2}
\]

\section*{C. Results}

Tables II, III and IV show the results of the cross validation testing on the geological resource data set with the NN and SQEXP kernels. The prediction errors and their standard deviation increase with block size as expected from Table I. This is because the support training data required for regressing at a test point is situated farther away. Increasing the test block size also results in reduced stratification as one fold of the cross validation may have e.g. 10,000 test points whereas another may have only 1000 points. With respect to the SE metric, the MTGP approach, using either kernel, performs better than both the equivalent GP approach (using its parameters for an exact comparison) and the independently optimized GP for each of the quantities, E1, E2 and E3. The GP models based on the NN kernel perform significantly better than those of the SQEXP kernel, with respect to all three metrics. This is in line with previous findings in [1]. In the case of the NN kernel, the MTGP model performs better than the derived and independent GP models in all three metrics. Not only is the SE lower, but the MTGP model produces lower uncertainty of prediction and lower NLP, suggestive of a more suited model than the equivalent GP or the independently optimized GP. These results clearly demonstrate that information integration across heterogeneous information sources is able to significantly improve individual models (predictions) of each of the types of information. In the case of the SQEXP kernel, the behavior of the VAR and NLP metrics suggests a poor MTGP (and consequently, equivalent GP) model. For a given high SE, the MTGP and GP models based on the SQEXP produce low
TABLE I
10 FOLD CROSS VALIDATION WITH BLOCK SAMPLING; 63667 POINTS IN DATA SET SPREAD OVER 3478.4 M X 1764.6 M X 345.9 M; BLOCK SIZES TESTED VS RELATIVE IMPLICATIONS ON RESULTS

| Block size | Method           | Number of points in fold with MIN test points | Number of points in fold with MAX test points | Comments on cross validation test |
|------------|------------------|-----------------------------------------------|-----------------------------------------------|----------------------------------|
| 87 x 45 x 9| MTGP             | 5807                                          | 6739                                          | Most Stratified cross validation |
| 174 x 89 x 18| GPI              | 5133                                          | 7549                                          | Less Stratified cross validation |
| 348 x 177 x 35| GP              | 4976                                          | 9662                                          | Least Stratified cross validation |

TABLE II
E1 CONCENTRATION ESTIMATION; 10 FOLD CROSS VALIDATION RESULTS USING BLOCK SAMPLING OF VARIOUS BLOCK SIZES; MULTI-TASK GP (MTGP) VS GP DERIVED FROM MTGP (GP) VS INDEPENDENTLY OPTIMIZED GP (GPI) USING NN & SQEXP KERNELS ON IDENTICAL TEST DATA

| Block size | Method   | SE (sqm) mean (std) | VAR (sqm) mean (std) | NLP mean (std) | SE (sqm) mean (std) | VAR (sqm) mean (std) | NLP mean (std) |
|------------|----------|---------------------|----------------------|----------------|---------------------|----------------------|-----------------|
| 87 x 45 x 9| MTGP     | 3.38 (22.92)        | 1.25 (5.06)          | 1.91 (2.56)    | 30.08 (70.98)       | 98.06 (25.34)        | 101.22 (13.42)  |
|            | GPI      | 85.88 (187.61)      | 58.60 (39.70)        | 3.71 (2.00)    | 265.25 (853.00)     | 1787.63 (244.59)     | 122.29 (23.44)  |
| 174 x 89 x 18| MTGP  | 14.60 (88.59)       | 10.97 (27.66)        | 2.24 (2.70)    | 52.96 (122.09)      | 70.62 (178.73)       | 34.69 (59.68)   |
|            | GPI      | 128.39 (261.10)     | 113.03 (88.51)       | 3.84 (1.60)    | 128.86 (244.59)     | 122.29 (23.44)       | 3.82 (0.96)     |
| 348 x 177 x 35| MTGP  | 7.42 (213.39)       | 64.36 (86.94)        | 4.06 (1.39)    | 114.47 (214.94)     | 101.93 (2801.29)     | 124.02 (94.50)  |
|            | GPI      | 189.14 (335.76)     | 199.86 (120.66)      | 3.98 (0.90)    | 190.92 (331.84)     | 155.24 (44.34)       | 24.18 (42.85)   |

TABLE III
E2 CONCENTRATION ESTIMATION; 10 FOLD CROSS VALIDATION RESULTS USING BLOCK SAMPLING OF VARIOUS BLOCK SIZES; MULTI-TASK GP (MTGP) VS GP DERIVED FROM MTGP (GP) VS INDEPENDENTLY OPTIMIZED GP (GPI) USING NN & SQEXP KERNELS ON IDENTICAL TEST DATA

| Block size | Method   | SE (sqm) mean (std) | VAR (sqm) mean (std) | NLP mean (std) | SE (sqm) mean (std) | VAR (sqm) mean (std) | NLP mean (std) |
|------------|----------|---------------------|----------------------|----------------|---------------------|----------------------|-----------------|
| 87 x 45 x 9| MTGP     | 8.04 (40.48)        | 6.22 (6.00)          | 2.37 (2.71)    | 211.05 (753.57)     | 1140.18 (8156.99)    | 103.63 (23.63)  |
|            | GPI      | 95.98 (274.95)      | 60.23 (41.14)        | 3.71 (2.27)    | 115.96 (301.71)     | 47.22 (22.50)        | 3.74 (1.22)     |
| 174 x 89 x 18| MTGP  | 21.49 (102.60)      | 15.88 (29.16)        | 2.66 (2.61)    | 402.24 (1144.48)    | 3510.60 (14423.99)   | 123.39 (41.47)  |
|            | GPI      | 142.62 (356.16)     | 123.42 (91.33)       | 3.88 (1.88)    | 164.86 (376.69)     | 312.01 (914.26)      | 3.92 (1.25)     |
| 348 x 177 x 35| MTGP  | 82.02 (233.18)      | 72.71 (861.66)       | 3.23 (2.04)    | 419.72 (1196.06)    | 8397.84 (20845.42)   | 194.54 (59.86)  |
|            | GPI      | 219.37 (484.76)     | 265.84 (239.30)      | 4.09 (1.71)    | 234.24 (459.31)     | 164.10 (633.41)      | 4.11 (1.17)     |

TABLE IV
E3 CONCENTRATION ESTIMATION; 10 FOLD CROSS VALIDATION RESULTS USING BLOCK SAMPLING OF VARIOUS BLOCK SIZES; MULTI-TASK GP (MTGP) VS GP DERIVED FROM MTGP (GP) VS INDEPENDENTLY OPTIMIZED GP (GPI) USING NN & SQEXP KERNELS ON IDENTICAL TEST DATA

| Block size | Method   | SE (sqm) mean (std) | VAR (sqm) mean (std) | NLP mean (std) | SE (sqm) mean (std) | VAR (sqm) mean (std) | NLP mean (std) |
|------------|----------|---------------------|----------------------|----------------|---------------------|----------------------|-----------------|
| 87 x 45 x 9| MTGP     | 2.87 (13.66)        | 2.79 (1.97)          | 1.88 (2.06)    | 9.19 (24.11)        | 35.35 (99.80)        | 21.21 (54.63)   |
|            | GPI      | 20.35 (55.84)       | 13.22 (8.54)         | 3.01 (2.45)    | 21.21 (54.63)       | 3.73 (7.16)          | 2.94 (1.17)     |
| 174 x 89 x 18| MTGP  | 6.63 (35.07)        | 6.00 (8.73)          | 2.17 (2.24)    | 14.36 (34.46)       | 73.85 (189.74)       | 12.06 (14.90)   |
|            | GPI      | 29.91 (81.55)       | 25.03 (19.65)        | 3.13 (2.01)    | 29.50 (72.11)       | 26.55 (6.71)         | 7.34 (14.06)    |
| 348 x 177 x 35| MTGP  | 22.76 (76.19)       | 23.57 (36.19)        | 3.64 (1.91)    | 27.46 (55.29)       | 96.50 (228.76)       | 34.15 (10.70)   |
|            | GPI      | 48.12 (113.43)      | 55.21 (53.40)        | 3.32 (1.82)    | 43.24 (93.69)       | 3.59 (10.17)         | 3.25 (1.35)     |

231
that performed both an exact and an independent comparison between MTGPs and GPs. Future work will concentrate on further benchmarking using other kernel, data-sets and on extending the approach to multi-data-set multi-task data fusion.

ACKNOWLEDGEMENTS

This work has been funded by the Rio Tinto Centre for Mine Automation.

REFERENCES

[1] S. Vasudevan, F. Ramos, E. Nettleton, and H. Durrant-Whyte, “Gaussian Process Modeling of Large Scale Terrain,” Journal of Field Robotics, vol. 26(10), pp. 812–840, 2009.
[2] C. E. Rasmussen and C. K. I. Williams, Gaussian Processes for Machine Learning. MIT Press, 2006.
[3] C. E. Rasmussen, “The Gaussian processes web site.” [Online]. Available: http://www.gaussianprocess.org/
[4] P. K. Kitanidis, Introduction to Geostatistics: Applications in Hydrogeology. Cambridge University Press, 1997.
[5] G. Matheron, “Principles of Geostatistics,” Economic Geology, vol. 58, pp. 1246–1266, 1963.
[6] M. El-Beltagy and W. Wright, “Gaussian processes for model fusion,” in International Conference on Artificial Neural Networks (ICANN), 2001.
[7] R. Murray-Smith and B. Pearlmutter, Deterministic and Statistical Methods in Machine Learning, LNBI 3635. Springer-Verlag, 2005, ch. Transformations of Gaussian Process priors, pp. 110–123.
[8] M. Girolami, “Bayesian data fusion with Gaussian process priors: An application to protein fold recognition,” in Workshop on Probabilistic Modeling and Machine Learning in Structural and Systems Biology (PMSB), 2006.
[9] S. Reece, S. Roberts, D. Nicholson, and C. Lloyd, “Determining intent using hard/soft data and Gaussian process classifiers,” in Proceedings of the 14th International Conference on Information Fusion (FUSION), 2011.
[10] S. Vasudevan, F. Ramos, E. Nettleton, and H. Durrant-Whyte, “Heteroscedastic Gaussian processes for data fusion in large scale terrain modeling,” in the International Conference for Robotics and Automation (ICRA), 2010.
[11] ——, “Large-scale terrain modeling from multiple sensors with dependent Gaussian processes,” in the proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Taipei, October 2010.
[12] ——, “Non-stationary dependent Gaussian processes for data fusion in large scale terrain modeling,” in the proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Shanghai, China, 2011.
[13] R. M. Neal, Bayesian Learning for Neural Networks, ser. Lecture Notes in Statistics 118. New York: Springer, 1996.
[14] C. K. I. Williams, “Computation with infinite neural networks,” Neural Computation, vol. 10(5), pp. 1203–1216, 1998.
[15] ——, “Prediction with Gaussian processes: From linear regression to linear prediction and beyond,” in Learning in Graphical Models, M. I. Jordan, Ed. Springer, 1998, pp. 599–622.
[16] K. Hornik, “Some new results on neural network approximation,” Neural Networks, vol. 6(8), pp. 1069–1072, 1993.
[17] A. Melkumyan and F. Ramos, “Multi-Kernel Gaussian Processes,” in the proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2011.
[18] D. Higdon, Quantitative Methods for Current Environmental Issues. Springer, 2002, ch. Space and Space-Time Modeling Using Process Convolutions, pp. 37–54.
[19] P. Boyle and M. Frean, “Dependent Gaussian processes,” in Advances in Neural Information Processing Systems 17, L. K. Saul, Y. Weiss, and L. Bottou, Eds. Cambridge, MA: MIT Press, 2004, pp. 217–224.
[20] E. Bonilla, K. M. Chai, and C. Williams, “Multi-task Gaussian process prediction,” in Advances in Neural Information Processing Systems 20, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. Cambridge, MA: MIT Press, 2007, pp. 153–160.
[21] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in International joint Conference on artificial intelligence, vol. 14, 1995, pp. 1137–1145.