Pressure matching for hydrocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments

Peter S. Craig
Michael Goldstein
Allan H. Seheult
James A. Smith

ABSTRACT In the oil industry, fluid flow models for reservoirs are usually too complex to be solved analytically and approximate numerical solutions must be obtained using a 'reservoir simulator', a complex computer program which takes as input descriptions of the reservoir geology. We describe a Bayes linear strategy for history matching; that is, seeking simulator inputs for which the outputs match closely to historical production. This approach, which only requires specification of means, variances and covariances, formally combines reservoir engineers' beliefs with data from fast approximations to the simulator. We present an account of our experiences in applying the strategy to match the pressure history of an active reservoir. The methodology is appropriate in a wide variety of applications involving inverse problems in computer experiments.

1 Introduction

This paper describes and illustrates, by means of a case study, a Bayes linear approach, incorporating expert knowledge, to matching historical hydrocarbon reservoir production using simulators of the reservoir. A reservoir simulator is a complex computer program for solving the time-dependent partial differential equations used to model oil, gas and water flow through an operating reservoir. It takes as input various aspects of reservoir geology and reservoir operation, including spatial distributions of rock properties and details of well locations and operating conditions. It returns as output various measures of production, mostly in the form of time series; in particular, oil, gas and water and pressures measured at the different wells. We are concerned with finding geology inputs for which simulator output closely matches observed historical production and, in particular, in this

study, we aim to match the historical record of pressure readings from an active reservoir.

Pressure matching is the first stage in a more general process of history matching, in which we seek a reservoir geology which yields as close a match as possible to the historical record for a wide variety of production variables. This search is carried out by running the simulator at different input settings until, hopefully, an acceptable match is found. The task is complex as the input and output spaces are very high dimensional, there may be many or no acceptable matches, and the simulator is expensive in CPU time for each run. In the case study, we considered 40 input variables and 77 output variables, and a simulation on our computer could take up to three days.

History matching is an example of a wide class of difficult practical problems involving high dimensional searches in complex computer experiments. Such computer experiments have the special feature that repeated runs at the same input settings always produce the same output, so that the classical concept of random error is not present; an overview of the field is given in [SWMW89]. We shall describe a quite general approach to pressure matching, which may be similarly applied to many related types of computer search problems. Our approach to prior specification is based around combining expert judgements of reservoir engineers with information gained from the analysis of experiments comprising a large number of runs on relatively fast simple versions of the simulator. For example, a reservoir engineer might judge that an output variable, such as the pressure at a particular well on a particular date, may depend primarily on just a few of the input variables, such as the permeabilities of neighbouring regions of the well. The engineer may also make assessments of magnitudes of measurement errors in the recorded production output. The simulator may be speeded up in various ways: in the case study, we coarsened the spatial grid employed to solve the fluid flow equations. The output from many relatively inexpensive runs of the faster simulator is used to build simple descriptions relating simulator output to simulator input. The two sources of information are combined into a specific structure describing prior means, variances and covariances for relationships between the simulator inputs and outputs, and these relationships are analysed and updated using Bayes linear methods given the outputs from relatively few runs of the 'full' simulator. Bayes linear methods (for an introduction, see [FG93]) only require specification of prior expectations, variances and covariances for the random quantities of interest, without the need for probability distributions, and they can be linearly updated given outputs on the full simulator. In this way, prior specification and subsequent analysis, including design computations for selecting inputs to the full simulator, are tractable, especially in contrast with frequentist and traditional Bayesian formulations.

Our intention is to describe both a general methodology and a case study which applies the methodology. A preliminary description of our strategy

in [CGSS96] gave various elements of the methodology with application to a small demonstration problem. In this paper, we apply the methodology to pressure matching for an actual reservoir. Among the differences that follow from the substantially larger scale of this application are that we need to consider (i) issues related to measurement errors in the recorded history; (ii) differences between the actual reservoir and the simulator; (iii) the detailed task of constructing the prior specification; and (iv) efficient ways to carry out the various computational steps in our procedures, involving high dimensional search and belief summarisation tasks.

The paper is set out as follows. In Section 2, we describe some of the background to the history matching problem. In Section 3, we give details of the particular reservoir that is the concern of this study. In Section 4, we discuss the objectives for the study. In Section 5, we outline the general Bayes linear strategy that we shall follow for matching pressures. In Section 6, we describe the prior formulation that we shall use for relating particular geology inputs to pressure outputs. In Section 7, we explain how we carried out the prior specification to quantify the various elements of the prior description. In Section 8, we develop the notion of an implausibility measure which we use as a basis for restricting the subset of input geologies which may yield acceptable matching outputs. In Section 9, we describe the diagnostics that we employ to monitor the various elements of our prior description. In Section 10, we explain how we choose which simulator runs to make in order to narrow our search for acceptable matches. In each section, we illustrate the general account with the relevant steps for our case study, and in Section 11 we describe further practical details of the case study. Finally, in Section 12, we make various concluding comments about the study.

2 History Matching

In the oil industry, efficient management and prediction of hydrocarbon production depends crucially on having a good fluid flow model of the reservoir under consideration. The model, which is a complex system of time-dependent, non-linear, partial differential equations describing the flow of oil, gas and water through the reservoir, is too difficult to be solved analytically. Instead, an approximate numerical solution is obtained using a 'reservoir simulator', a computer code which takes as input physical descriptions of the reservoir, such as geology, geometry, faults, well placements, porosities and permeabilities, and produces as outputs time series of oil, gas and water production and pressures at the wells in the reservoir. The corresponding historical time series are mostly unequally spaced and the times at which the different series are recorded often do not correspond. The eventual aim is to find a setting of the input geology which results in

a simulator run with outputs which match as closely as possible the corresponding reservoir history. This process is termed history matching and is described in [MD90]. In the case study, we try to match recorded pressures by varying permeabilities in the geologically distinct subregions that make up the reservoir and by varying transmissibilities of faults which geologists believe may exist at certain locations.

There are several reasons for carrying out a history match. First, we may hope to predict future reservoir performance, using the given simulator. Secondly, we may intend to use the simulator as an aid to decision making, for example to help decide where to sink further wells. Thirdly, we may want to make inferences about the physical composition of the reservoir which may be used in a more general context, for example as inputs into a more detailed simulation study of the reservoir. Of most immediate relevance for the approach that we shall develop are the following two distinctions:

1. Is the aim to find a single good history match, as is often the current practice, or is it to seek to identify the class of all possible history matches, subject to some general criterion of match quality?

2. Is the aim to find history matches which are predictive for future reservoir performance, or is it to find matches which are informative for the composition of the reservoir?

We will discuss the consequences of these methodological distinctions in our development of the case study.

In practice, this inverse problem, namely choosing an appropriate input geology and perhaps making structural changes to the model so that the simulator output matches historical time series, is tackled by reservoir engineers on a trial and error basis using all of their technical knowledge, experience and specific understandings of the particular reservoir under study. At each stage, values are chosen for input geology settings, the simulator is run, the production output is scrutinised and compared with historical production, and new settings for the geology inputs are suggested. This process is time consuming for two reasons. First, simulator run-time is typically many CPU hours (between 1 and 3 days in our case study), so that it is impractical to make large numbers of runs. Secondly, the output from each run requires careful scrutiny in order to select the next choice of input values at which to run the simulator. In many cases, a fully satisfactory match is not obtained and usually it is hard to judge whether this was due to a problem with the underlying model or due to an inadequate search over the space of possible inputs. It is hoped that, by formalising aspects of the problem, (i) more efficient choices may be made for runs on the simulator; (ii) the efforts of the reservoir engineer may be better spent on those aspects of the study which genuinely require expert intervention; and (iii) a rather more informative analysis can be carried out, involving consideration of the class of possible matches and associated uncertainties.

While we discuss these issues in the context of history matching, they are similarly relevant for the analysis of any large computer model.

3 Case Study

This case study concerns a large, primarily gas-producing reservoir, one with which our industrial collaborators Scientific Software-Intercomp (SSI) have twice been involved. For reasons of commercial confidentiality, the identity and location of the reservoir and certain other details cannot be disclosed.

The reservoir covers an area of about 30km by 20km and is about 60 metres thick. The main productive reservoir units are two sandy layers, a top layer of clean marine sand, and a more shaley bottom layer.

The reservoir comprises four fields. One field is mainly onshore with a small part under shallow coastal waters, and the other three are offshore in progressively deeper water. Gas production has to date been solely from the onshore field, due mainly to the relative costs and ease of drilling. Historical performance suggests that the reservoir fields effectively behave as one unit, pressure communication between individual fields being via the hydrocarbon zones and/or common aquifers.

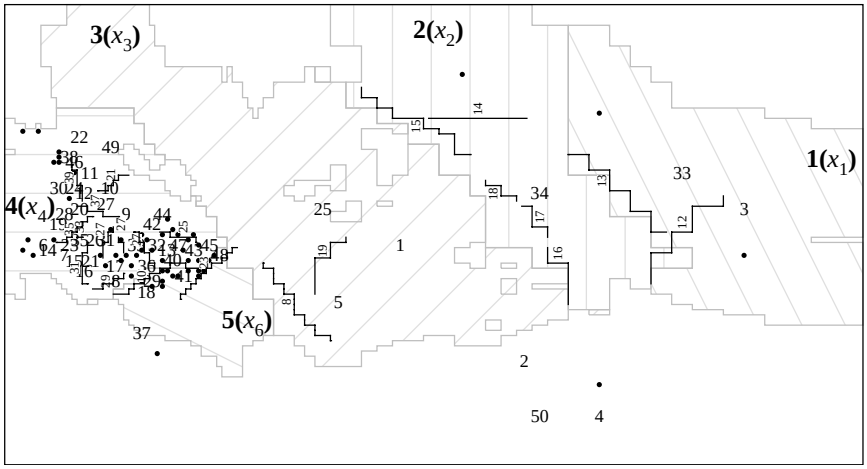
A reservoir description developed by SSI in 1993/94, covering the production period March 1972 to June 1993, formed the basis of the three-dimensional simulation model used in this case study.

3.1 Simulation Model Description

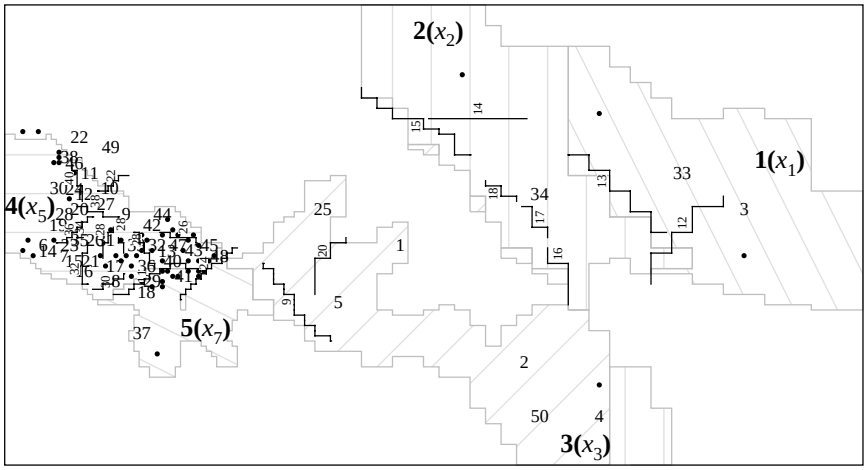
The reservoir description, which includes reservoir structure and geometry, spatial distribution of porosity and fault patterns, was provided by SSI. Figures 1 and 2 show the extent of the two reservoir layers, and the approximate location of wells and possible faults.

The *initial* spatial distribution of absolute permeability for this reservoir was derived from the corresponding spatial distribution of porosity supplied by the client oil companies. In the case study, the reservoir was divided into seven sub-regions. For each sub-region a single multiplicative factor, varying between 10^{-1} and 10^{1}, was used to modify the initial map. Permeabilities in the directions of the chosen horizontal orthogonal axes were taken to be the same, while vertical permeability was taken to be 1% of horizontal permeability. Porosity and other 'parameter' values supplied by SSI, such as capillary pressures, were kept fixed throughout the case study.

Early simulation studies by SSI utilised a two layer model corresponding to the two sandy units. A five layer refinement with these two layers vertically subdivided was used by SSI in the later stages of their study. In our study of the same field over the same production period, we used this

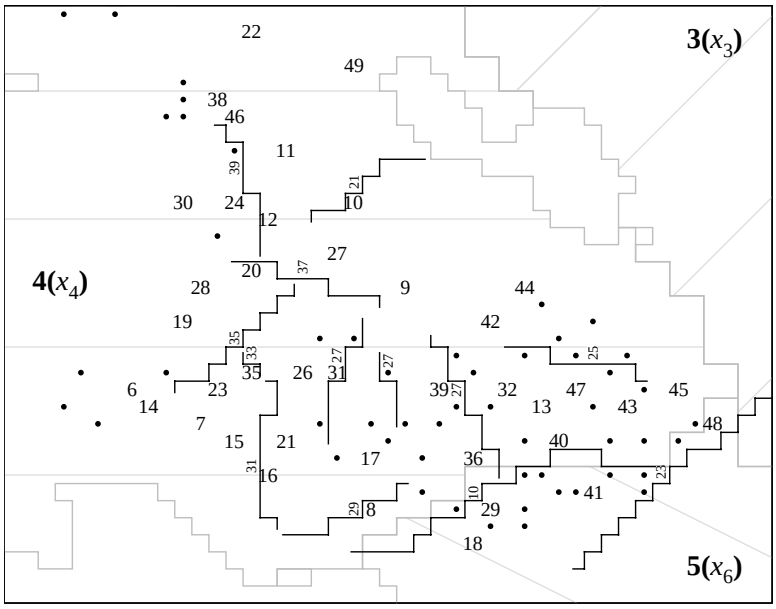


Top layer of reservoir

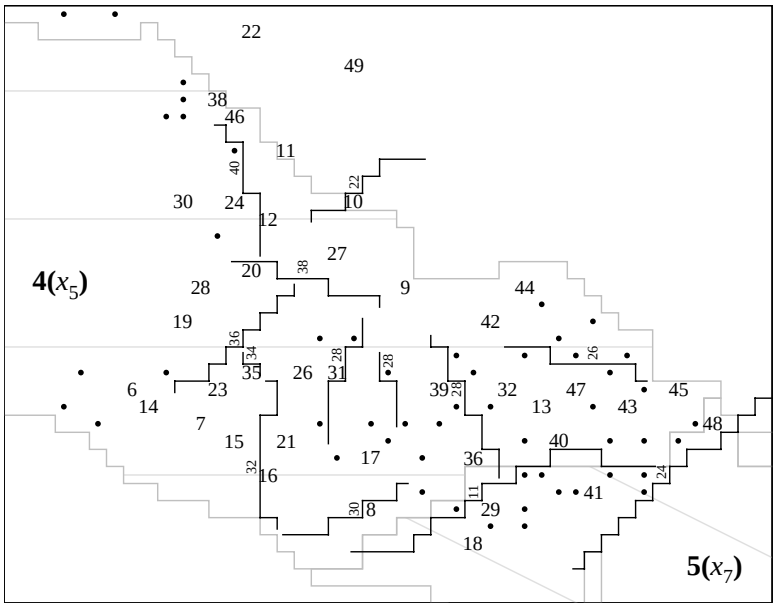


Bottom layer of reservoir

FIGURE 1. Maps of the layers of the reservoir (magnified maps of the western end of the reservoir are shown in Figure 2). Regions are indicated by regular hatching and labelled by bold-face numbers. The corresponding simulator input variable, acting as permeability multiplier for the region, is shown in parentheses after the region number . White space areas are not part of the reservoir. Possible faults are indicated by thicker lines. A number on its side beside a fault indicates the variable controlling the fault's transmissibility. Wells with measurements used in the study are located at the centre of the corresponding well number. Other wells are indicated by dots.



Top layer of reservoir



Bottom layer of reservoir

FIGURE 2. Magnified maps of the western end of the reservoir. For details of the meaning of symbols, see Figure 1.

same five layer model as our full simulator, and the fast simulator, which we ran many times to develop our prior specification, was a coarser grid version of SSI's original two-layer simulation model.

SSI's original reservoir simulator was based on an 80×60 non-uniform cartesian grid on each of the two layers, shown in the top panel of Figure 3. To better approximate the production of the reservoir, SSI's refinement to the two layer model slices the top layer into three layers and the bottom into two layers. The refinement offers a more realistic model of fluid flow. The run time of the five-layer simulator is between one and three days CPU time on our SPARC10.

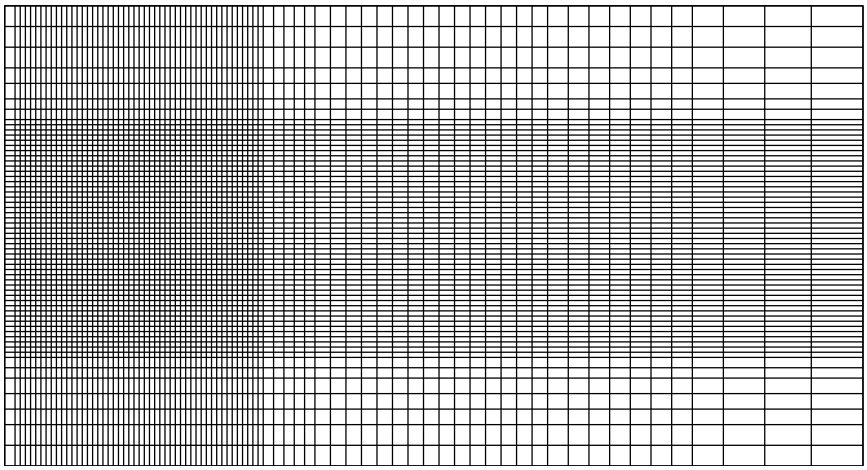
For the case study, our fast simulator was obtained by coarsening the simulation grid in SSI's two layer model. The coarsened grid is shown as the lower panel of Figure 3. The coarsening was chosen to keep each well in its own cell, not to alter the regions and faults too much, and to keep the relative sizes of neighbouring cells reasonably close, as large discrepancies in neighbouring cell sizes can cause simulation difficulties. The coarsening reduced the run time to about twelve minutes, a reduction by a factor of more than 100 by comparison with the full simulator. It should be noted that in the coarsening process a well's position may change to locate it in the centre of its cell, resulting in possible bias between the two simulators.

3.2 The Production History

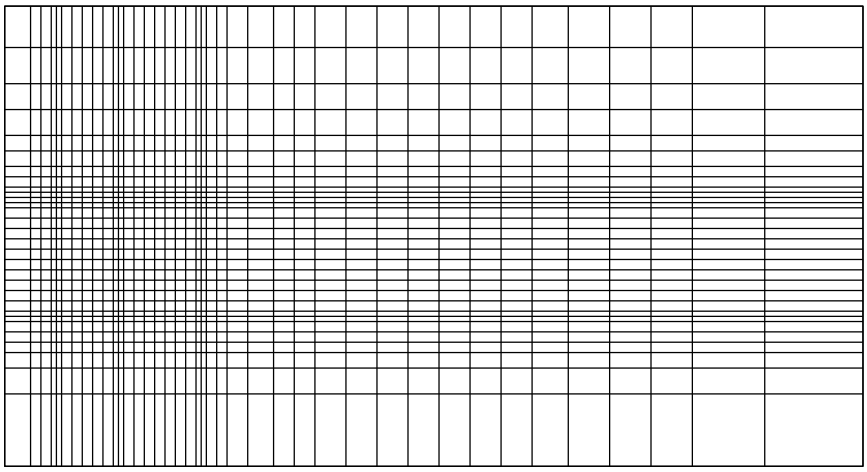
The oil companies working the field have drilled a total of 109 wells, of which 89 are production wells, and the rest are mostly offshore pressure test wells. Production history is available on an irregular monthly basis from March 1972 to June 1993. For the pressure match, there were a total of 221 pressure readings taken on 68 wells, including all of the non-production wells. There are no more than 9 pressure readings for any one well. As well as pressure, there are production series for gas rate, oil rate, gas/oil ratio and water, the pressure readings being the sparsest of all the series.

3.3 Matching Pressure History

This case study reports the first stage of history matching, which, in accord with standard practice, is to try to match the pressure data. To avoid problems with setting up the simulator, we chose just the 77 pressure readings for which we had both pressure history *and* output from the simulator. Some of these pressure readings were recorded at different times at the same well. In all, we used pressure readings from 50 wells, indicated by numbers in Figures 1 and 2. Table 1.1 summarises this history: column 2 indicates the well number, used to mark the location of a well in Figures 1 and 2; column 3 gives the times of the pressure readings in months from the start of the study; column 4 gives the reservoir region containing the



Original Flow Simulation Grid



Coarsened Grid

FIGURE 3. Grids used for full and fast simulators of reservoir.

well; the formation of the pressure readings into the twelve groups in column 5 is based on their well location and time of measurement, and will be explained in more detail in Section 7.4; and the label y_{H_i} in the last column refers to the i-th historical pressure reading.

During initial consultation with the reservoir engineer who last worked on the study at SSI, it was also decided to divide the reservoir into five regions which we refer to as regions 1 to 5, shown in Figures 1 and 2. Regions 4 and 5 were each split into top and bottom layers, giving seven sub-regions in all. Regions 1 to 3 were not split as there is insufficient historical data in this part of the reservoir. In the history matching exercise, the permeability in each of the seven sub-regions was altered by multiplying its initial permeability map by a single multiplier. As mentioned above, these initial permeability maps were derived from the porosities supplied by the client oil companies. Figures 1 and 2 show which of the permeability multipliers, $x_1, \ldots, x_7$, applies to which sub-region.

The geologists believed there were 20 active faults in the reservoir, and their location and extent were supplied. As well as altering the permeabilities of the sub-regions, it was decided to alter the fault transmissibilities. A fault transmissibility multiplier is between *zero* (a sealing fault with no flow across it possible) and *one* (no effect on flow). If a fault lies in any of the right hand regions (1, 2 or 3) it is given a common fault transmissibility multiplier in both layers, whereas the transmissibility of a fault in one of the left hand regions (4 or 5) can be varied independently in both layers. Of the 20 active faults in the reservoir, seven are in regions 1, 2 or 3 and thirteen in regions 4 or 5, giving a total of 33 multipliers $x_8, \ldots, x_{40}$. In Figures 1 and 2, the input variable controlling the transmissibility of a fault is indicated, by a number on its side, beside the fault.

4 Objectives of the Study

There are various motivations for carrying out the history match for the given reservoir, from specific questions concerned with learning about the geology in the right-hand region, to more general questions concerning the overall composition of the reservoir. One important question the drilling company wished to answer, but not considered here, was whether the gas and oil in the offshore fields could be recovered solely from the current onshore wells. While, for a full history match, we would compare the simulator output with all of the recorded history, for the purposes of this study we will restrict attention to seeking those collections of input values, specifically values for regional permeability multipliers and fault transmissibility multipliers, which achieve an acceptable match for all the pressures in Table 1.1.

As the pressure match is the first stage in a larger matching process, it is

i	Well	Time	Region	Group	y_{Hi}	i	Well	Time	Region	Group	y_{Hi}
1	1	116	3	8	2292	40	20	188	4	3	1711
2	2	130	3	9	2335	41	9	188	4	3	1719
3	3	132	1	9	2364	42	12	188	4	3	1707
4	4	132	3	9	2350	43	13	188	4	3	1727
5	5	134	3	9	2234	44	29	188	5	6	1695
6	6	138	4	1	1723	45	30	188	4	3	1707
7	7	138	4	1	1769	46	31	188	4	3	1362
8	8	138	4	1	1880	47	32	188	4	3	1700
9	9	138	4	1	1906	48	21	188	4	3	1668
10	10	138	4	1	1913	49	14	188	4	3	1613
11	11	138	4	1	1925	50	16	188	4	3	1682
12	12	138	4	1	1941	51	23	188	4	3	1624
13	13	138	4	1	1881	52	17	188	4	3	1726
14	14	138	4	1	1711	53	24	188	4	3	1711
15	15	138	4	1	1784	54	33	188	1	11	2263
16	16	138	4	1	1784	55	34	190	2	11	2154
17	17	138	4	1	1885	56	19	251	4	4	1376
18	18	148	5	5	1797	57	35	251	4	4	1407
19	19	148	4	2	1736	58	26	251	4	4	1436
20	8	148	4	2	1839	59	36	251	4	4	1501
21	20	148	4	2	1785	60	37	251	5	7	1545
22	10	148	4	2	1864	61	10	251	4	4	1458
23	13	148	4	2	1849	62	38	251	4	4	1443
24	21	148	4	2	1752	63	13	251	4	4	1455
25	14	148	4	2	1706	64	39	251	4	4	1398
26	22	148	4	2	1851	65	30	251	4	4	1423
27	15	148	4	2	1682	66	40	251	4	4	1424
28	16	148	4	2	1668	67	41	251	5	7	1475
29	23	148	4	2	1682	68	42	251	4	4	1664
30	17	148	4	2	1769	69	43	251	4	4	1513
31	24	148	4	2	1784	70	44	251	4	4	1501
32	25	166	3	10	2016	71	45	251	4	4	1498
33	26	186	4	3	1651	72	46	251	4	4	1434
34	7	188	4	3	1662	73	47	251	4	4	1466
35	18	188	5	6	1711	74	48	251	5	4	1553
36	19	188	4	3	1635	75	49	251	4	4	1433
37	8	188	4	3	1701	76	14	251	4	4	1369
38	27	188	4	3	1716	77	50	254	3	12	2016
39	28	188	4	3	1716						

TABLE 1.1. Historical data: well number (see Figures 1 and 2); time in months from start of simulation; region containing well; observation group as in Section 7.4; and historical pressure readings y_{Hi}, in lb/in^2.

important to identify the collection of possible geologies which may produce adequate pressure matches. In particular, it is useful to identify which of the input geology variables considered have little effect in determining the pressure match, as these may be varied in subsequent matching without affecting the match on pressure. More generally, the collection of potential matches may be used to capture uncertainty about predictions for future reservoir behaviour by running the simulator at different reservoir geologies which achieve acceptable match quality.

There are three views that we may take in assessing match quality. First, we may simply consider the difference between the observed history and the production output of the simulator. Secondly, as the observed history consists of true production values observed with error, we may seek to match the unobserved true production. Thirdly, we may consider that our goal is to identify the physical composition of the reservoir, in which case we must also incorporate into our comparisons the possible differences between the reservoir and the simulator. Although the first of these alternatives is usual practice, it seems more natural to prefer the second, incorporating explicit assessment of errors in measurement. The choice between the second and third view is less clear, however. If our intention is to predict future performance using the given simulator, then arguably it may be most efficient to choose the best predictor of true production from the simulator. However, if the results of the match are to be used for other inferential purposes, and in particular if the estimated geology that is obtained by the match is to be used outside the context of the given simulation, then arguably we should incorporate assessments of the differences between the reservoir and the simulator into our matching procedures.

In this study, we shall mainly concentrate on matching true production, largely ignoring the differences between reservoir and simulator and assessing the quality of the match in terms of judgements, about measurement errors in the pressure readings, that were provided by the reservoir engineer. Essentially, ignoring the differences between reservoir and simulator, avoids introducing an extra, and rather subtle, layer of modelling which would make it difficult, within the restrictions of the current account, to assess the success of our matching procedure. However, we could incorporate such differences into our procedures in a similar way to measurement error, and we shall indicate where such differences enter the analysis.

5 Bayes Linear Strategies for History Matching

In [CGSS96], some general strategies were outlined for tackling computer experiments such as history matching. In this paper we further develop these strategies and describe their application to a particular reservoir study.

The formal problem is as follows. At each run of the simulator, with an input vector $x = (x_1, \ldots, x_d)$ of reservoir properties, the simulator produces an output vector $y = (y_1, \ldots, y_p)$. In the case study, $d = 40$, $p = 77$, and each y_i is the pressure at a particular well at a particular time. We write y as $y(x)$ when we need to emphasise the dependence of y on x. We have a region R of values of x which are deemed by a geologist who has studied the reservoir to be, a priori, plausible as values corresponding to the actual geology of the reservoir. The history of recorded pressure outputs at the various wells is represented by a vector y_H of observations. Our objective is to identify the collection R_M of input choices $x \in R$ for which the output $y(x)$ is sufficiently 'close' to y_H (according to some scale that we must devise) to suggest that x merits further consideration as a setting of the geology inputs which we may use for predicting future behaviour of the reservoir. The set R_M therefore depends on how we measure match quality, and we may use a variety of criteria to identify different choices for this set.

Our formulation is in some ways different to the usual view of history matching in the oil industry, where often the task, as specified by contract, is to find a single value of x for which almost all the elements of $y(x)$ are within a specified percentage of the values of y_H. We are both more cautious, in that we consider that the analysis of the simulator can only identify a class of candidates for the reservoir geology, and more ambitious, in that our procedures aim to identify this class of matches. Therefore, we do not specify and update quantitative beliefs for the 'true but unknown' value of x, but instead specify and update beliefs about the value of $y(x)$ for each x, from which we can update beliefs about the collection R_M. One of the main reasons why history matching has usually been treated as the problem of identifying a single match is the difficulty of identifying classes of matches without using some type of Bayesian argument such as we shall develop in this case study. How the class of matches should be used is beyond the scope of this paper. At the least, predictions of reservoir performance may be compared over a variety of elements of R_M, to see whether the predictive differences within the set are substantial.

The difficulty in identifying R_M is that $y(x)$ is an extremely complex, high-dimensional function, which can take a very long time (in our case study, several days) to evaluate for a single choice of inputs. Therefore, it is appropriate, for each x, to view the value $y(x)$ as an unknown quantity, which may only be evaluated at high cost in CPU time. However, as x changes, values of $y(x)$ are highly correlated, so that each time we evaluate one such vector, we may modify our beliefs about all other such vectors. Therefore, informally our task is as follows. We specify joint prior beliefs for all of the values $y(x)$, for $x \in R$, and progressively update these beliefs as we make runs on the simulator. At each stage, our beliefs about the value of each $y(x)$ allow us to identify those x for which we currently believe that $y(x)$ is 'likely to be close' to y_H, from which we may choose our current

candidates for the collection R_M.

We have at least three types of problem. First, we have a modelling problem, in that we must find a way to develop joint prior beliefs about the values of all of the outputs for all choices of inputs. As we can only make a relatively small number of evaluations of the simulator, these prior beliefs must be sufficiently detailed and reliable to allow us to extrapolate our beliefs over the values of $y(x)$ for x varying over the whole of the region R, while still being of a sufficiently simple form both to allow for meaningful elicitation and also to result in a tractable analysis. Secondly, we have a design problem, in that we must at each stage be able to identify at which values of the inputs to run the simulator, so as to maximise the amount of information that we will gain about R_M. Thirdly, in addition to the general problem of updating such a large collection of beliefs, we have a substantial practical problem of inversion, which arises due to the high dimensional nature of the simulator. Even if the function $y(x)$ could be computed very cheaply for all x, there would still be a substantial numerical inversion problem in identifying the region R_M, and this difficulty is compounded by the substantial uncertainty as to the value of $y(x)$ for all values of x at which the simulator has not been evaluated.

Each of these problems is very difficult to tackle within a full Bayes formalism: prior specification is difficult due to the extreme level of detail required for the joint prior distribution; full Bayes design calculations are notoriously computer intensive, even for much more straightforward problems; and probabilistic analysis and inversion is technically difficult for high dimensional joint distributions.

For these reasons, we choose an approach which is based around the ideas of Bayes linear modelling and adjustment of beliefs. An overview of Bayes linear methodology, with applications, is given in [FG93]. The Bayes linear approach is similar in spirit to the Bayes approach but, rather than working with full prior beliefs, we restrict attention to the specification of prior means, variances and covariances between all quantities of interest. We update our beliefs by linear fitting on the data. In comparison with the full Bayes analysis, in the Bayes linear approach it is therefore easier to make the required prior specifications and also more straightforward to update beliefs, both of which are crucial simplifications when analysing very high dimensional sequential search problems.

The simplest form of the Bayes linear approach is as follows. We have two vectors B and D. We specify directly, as primitive quantities, the prior mean vectors, $\mathrm{E}[B]$, $\mathrm{E}[D]$, and prior variance and covariance matrices, $\mathrm{Var}[B]$, $\mathrm{Var}[D]$ and $\mathrm{Cov}[B, D]$. The adjusted expectation $\mathrm{E}_D[B]$ and the adjusted variance $\mathrm{Var}_D[B]$ for B, having observed D, are given by

$$\mathrm{E}_D[B] = \mathrm{E}[B] + \mathrm{Cov}[B, D]\mathrm{Var}[D]^{-1}(D - \mathrm{E}[D]) \quad (1.1)$$

$$\mathrm{Var}_D[B] = \mathrm{Var}[B] - \mathrm{Cov}[B, D]\mathrm{Var}[D]^{-1}\mathrm{Cov}[D, B] \quad (1.2)$$

If Var[D] is not invertible, then we use the Moore-Penrose generalised inverse in the above equations. Adjusted expectations may be viewed as simple approximations to full Bayes conditional expectations, which are exact in certain important special cases such as joint normality. An alternative interpretation of Bayes linear analyses, as those inferences for posterior judgements which may be justified under partial prior specifications, is given in [Gol96]. While there are important foundational reasons for using the Bayes linear approach, our choice in this case study is made on pragmatic grounds. The full Bayes approach is too difficult to implement and our hope is that the Bayes linear approach will capture enough of the uncertainty in the problem to allow us to develop tractable and effective strategies for history matching.

Our general strategy, which is applicable to a wide variety of inverse problems for large scale computer models, may be summarised in the following steps, which provide an overview of the methodology that we describe and apply in this study.

1. We elicit aspects of qualitative and quantitative prior beliefs by discussion with a reservoir engineer. We only quantify prior means, variances and covariances. We need to elicit two types of beliefs. First, we must consider the relationship between the observed history y_H and the output from the simulator. Secondly, we must consider the relationship between input and output for the simulator. In principle, the elicitation should proceed as follows, though in practice we may need to approximate some of the steps.

   (a) We must consider how closely we should expect to be able to match the observed history y_H. There are two reasons why we may not achieve a close match. First, the observations y_H may be considered to consist of a vector of true pressure values, y_T, observed with measurement errors, whose prior variances must be quantified. Secondly, there may be systematic differences between the simulator and the reservoir, and means, variances and covariances for such differences should be assessed. In this study, more attention was paid to representing measurement errors than to representing differences between the reservoir and simulator, but we shall discuss both types of representation and the issues involved in relating each of these assessments to match quality.

   (b) For each $x, x' \in R$, we must elicit a prior mean vector and variance matrix for $y(x)$ and a covariance matrix between $y(x)$ and $y(x')$. These mean and variance specifications are functions of the input vector. Given the high dimensionality of the problem, and the need for detailed investigation of the mean and variance surfaces, we construct this prior specification by first eliciting

simple qualitative descriptions which focus on the most important relationships between the inputs and the outputs. We therefore seek to identify, for each component $y_i(x)$ of $y(x)$, a small subset of components of x, termed the active collection of inputs for y_i, with the property that our beliefs about changes in the value of $y_i(x)$ resulting from changes in x are almost entirely explained by the changes in the values of the active variables for y_i. In the study, we found that, for each output quantity, the reservoir engineer was willing to restrict attention to a collection of three active variables, which resulted in enormous simplifications in the subsequent analyses. For each output, we must therefore identify the active inputs qualitatively, and then quantify prior beliefs about the magnitudes of the effects of each active input, which we do by considering the following type of assessment: 'if the value of a particular input, x_j say, changes by one unit, while holding all other inputs fixed at some central values, assess means and variances for the change in each component of $y(x)$ for which x_j is an active variable.'

(c) There is a third potentially useful source of prior beliefs that may be relevant to this problem, namely a prior specification of beliefs about the physical geology. However, we did not have access to the geologist who had been familiar with this reservoir, and the reservoir engineer was unwilling to express such beliefs. It was also his opinion that the geologist was unlikely to have strong prior opinions.

2. Identifying active variables for each output and quantifying their effects is a large, challenging and unfamiliar elicitation task. Therefore, we find it helpful to construct fast approximations to the original simulator, which can provide additional information to guide our prior quantification. We make many runs on the fast simulators, for inputs chosen within region R. In the case study, we could perform roughly 200 runs on the fast simulator for the time taken by one run of the full simulator. By fitting linear models to the data from the runs on the fast simulator, we may quickly form a fairly detailed, albeit provisional, qualitative and quantitative picture of the relationships between the input and the output variables. Our aim, as in the prior elicitation, is to construct simple descriptions, where we select a small number of geology input variables which are most important in explaining the variation for each pressure output variable, and then to build a prior description which expresses these relationships. The prior specification process that was followed for the case study resulted in choosing, for each output quantity, a maximum of three active input quantities and fitting a quadratic surface in the three variables plus a spatially correlated residual structure.

3. We compare the choice of active variables from the prior elicitation and from the data analysis on the fast simulator, and combine the information from the two sources using a variety of informal judgements and formal methods based on quantities such as elicited beliefs concerning the differences between the two simulators.

   The overall result of our elicitation is (a) a specification of the relationship between the observed history and the output of the full simulator and (b) a prior description which specifies prior means, variances and covariances between all elements $y_i(x)$ and $y_j(x')$ on the full simulator, based on the selection of active variables for each component. The functional form of this latter description is based on small subsets of the elements of x and contains certain unknown random quantities (the coefficients of the linear and quadratic terms in the active variables) about which we may update beliefs as we make observations on the full simulator.

4. Using the prior description that we have created, for each $x \in R$, we now have an expectation vector and variance matrix for the difference $y(x) - y_T$. Informally, for any x for which $\mathrm{E}[y(x) - y_T]$ is a large number of standard deviations away from zero, we judge that it is unlikely that the actual value $y(x)$ would provide an acceptable pressure match. We therefore construct informal 'implausibility' measures based on the standardised distance from $\mathrm{E}[y(x) - y_T]$ to zero, which we use to identify and eliminate parts of the region R which we judge to be implausible as providing potential history matches. These implausibility measures are updated when we modify our expectation vectors and variance matrices as we obtain new information by making evaluations on the full simulator.

5. We now make a series of runs on the full reservoir simulator, sequentially selecting input vectors $x_{[1]}, x_{[2]}, \ldots$, and evaluating for each, the output vector $y(x_{[j]})$. Each such evaluation, $y(x_{[j]})$, may be used as data to reduce uncertainty for the value of $y(x)$ for each remaining x, using (1.1) and (1.2), with B equal to $y(x)$ and D being the vector of all previous evaluations $y(x_{[j]})$. We are not interested in learning about the value of any $y(x)$ for which, according to our current beliefs, the standardised value of $\mathrm{E}_D[(y(x) - y_T)]$ is far from zero. At each stage, we therefore identify the collection of simulator inputs which have high plausibility as potential pressure matches, and choose the value of x at which to evaluate $y(x)$ as that choice of input which most reduces our uncertainties about the values of $y(x)$ for this collection. There are various ways in which these choices may be made, depending on whether we are trying to find a single good pressure match, or whether we are trying to identify the collection of possible matches. We largely follow the latter alternative and so make sequen-

tial choices intended to eliminate regions of input space from further consideration.

6. After each run of the simulator, we proceed as follows.

   (a) We evaluate various diagnostics to check the reliability of our prior description. These diagnostics compare the observed values of the output $y(x)$ with our current expectation and variance for $y(x)$. Consistently poor performance of these diagnostics may simply suggest an automatic modification, such as inflating some of the variances in our description, or might require expert intervention to rebuild various aspects of the belief specification. We might make further calculations to help us to improve the belief specification. For example, we might learn further about the relationship between the fast and the full simulator, both by assessing the variance of the difference between the outputs for the two simulators given identical inputs, and also by exploring whether there might be systematic differences between the two simulators. In the case study, we did not develop such additional analyses, mainly due to time constraints.

   (b) Provided that the diagnostics are satisfactory, the current expectations and variances for each $y(x)$ are updated using Bayes linear fitting. Our various implausibility measures are re-evaluated as a basis for the next choice of simulator run.

7. As we continue to make runs on the simulator, we expect the region of geology inputs which might plausibly yield an adequate match to shrink in volume. When the region has been sufficiently reduced, which we assess by visual inspection of our implausibility measures, a reduced region is formally selected. Our analysis may suggest that this region should contain geologies which were not within the region originally selected as most plausible by the geologist. We then repeat step 2 of our strategy, making many runs on the fast version of the simulator over the reduced region, and recreating our prior description. This process allows new choices of input geology variables to become active within our revised prior descriptions and also allows us to re-estimate the effects of input variables which remain active in the restricted sub-region. If the new region consists of several disconnected sub-regions, we would consider constructing separate descriptions for each sub-region.

8. We now repeat steps 3 to 6 on the restricted region, until our plausibility measures again suggest that the region may be further restricted by repeating step 7.

9. This process is iterated until, subject to constraints of time and model and measurement accuracy, we have identified those collections of geology inputs which yield an acceptable pressure match. At this stage, we may seek to identify geology inputs within our reduced region which we consider to give particularly close pressure matches, as an indication of the overall degree of success of the matching process.

6 Prior Formulation

We now describe the various components of the formal description of prior beliefs that we use in the case study. The description has two parts. First, we must represent our beliefs concerning the history of recorded pressure values at the various wells. Secondly, we must represent our beliefs about the outputs $y(x)$.

The observed values of the pressures are represented by a vector y_H. We consider the observations to be the sum of the actual pressure values at the wells, which we denote by the vector y_T, and a vector of measurement errors y_E, namely

$$y_H = y_T + y_E \tag{1.3}$$

where y_T and y_E are uncorrelated prior to observing y_H. The vector y_E has zero prior mean and usually, but not necessarily, its components are uncorrelated. Throughout we consider y_E to be uncorrelated with all quantities $y(x)$, ignoring for simplicity any correlations that may arise from beliefs about R_M.

To express the differences between the reservoir and the simulator, we may introduce the further relations

$$y_T = y_C + y_D \tag{1.4}$$

where y_C is the simulator output that we would obtain if we had perfect information about the reservoir and used this to choose the corresponding values of the geology inputs to the simulator; and y_D represents the difference between the simulator and the reservoir. However, y_C is not fully operationally defined by the above description, as the simulator is a simplification and abstraction of the reservoir. Thus, we might use a further level of modelling to link the actual reservoir geology with the representation in the simulator. However, this is an unnecessary complication for this account, and equation (1.4) is an adequate representation for many practical purposes. In our specification, y_C, y_D and y_E are mutually uncorrelated. The components of y_D may be highly correlated; for, if the simulator makes a poor approximation to a particular aspect of production history, then it is likely that there will be a similarly poor approximation to related aspects of that history.

We may have various objectives when history matching. The usual formulation within the oil industry is to try to find values x for which $y(x)$ is in some sense close to y_H. Intuitively, it is more meaningful to seek values x which are close to y_T. Of course, y_T is unobservable, but the practical effect is that we will be more concerned to match closely those elements of y_H which we judge to be accurately measured than those which we judge to be poorly measured, as represented by the prior variances for the components of y_E.

A further alternative is to seek to match y_C, with the further practical effect that we will be more concerned to match those elements of y_H which correspond to aspects of reservoir production which we consider to be well represented by the simulator. For certain purposes, matching y_C seems to be a sensible objective, for example if we want to use the match to make generalised inferences about the geology of the reservoir. However, if our aim is to choose an input vector x for which the simulator output provides a good forecast of future performance of the reservoir, then the choice is less clear. The view of the reservoir engineers whom we consulted was that they sought inputs which were able to reproduce the previous reservoir performance, rather than representing the actual reservoir.

We therefore chose, for the present study, to seek choices of inputs for matching y_T. Matching y_C would proceed in a formally similar way, but would rely on a careful elicitation of beliefs over the elements of (1.4).

We now describe how we express prior beliefs about the value of the output vector $y(x)$ for general x. We must specify an expectation vector and variance matrix for each $y(x)$ and a covariance matrix for each pair $y(x)$ and $y(x')$. In principle, these quantities are functions of all of the elements of x, but, for each component $y_i(x)$, certain elements of x will be the most important in determining the mean and variance structure. This is an important consideration in making the prior elicitation and the resulting analysis tractable. In our prior formulation, we therefore construct beliefs about each $y_i(x)$ which are mainly determined by a small number of such components of x. For each component $y_i(x)$, we subjectively identify a subset, x_i^*, of the components of x, termed the active variables for y_i, where, usually, x_i^* is a small subset of x. In our case study, any x_i^* has at most three components. These quantities are chosen to be the elements within the input vector which are most important in accounting for our current beliefs about variation in y_i as x varies over the region of geology values which are currently under consideration for providing a potential history match. The choice of active variables is an exercise in qualitative prior elicitation, and in Section 7, we will describe how, in practice, this elicitation is made. As we proceed with the analysis, we may change the choice of active variables several times, and we will describe how and why the collection may change in our description of the progress of the case study.

Having selected the active variables, we express our prior beliefs about

the relation between y_i and x as an equation of the form

$$y_i(x) = \alpha_i + \sum_j \beta_{ij} g_{ij}(x_i^*) + \epsilon_i(x_i^*) + \delta_i(x) \tag{1.5}$$

where the $g_{ij}(x_i^*)$ are known simple functions of x_i^*, and ϵ and δ are discrepancy terms from the simple linear fit, with zero expectations and with variance structures which we will discuss below. Equation (1.5) is not intended to represent a physical model for the relationships between the inputs and outputs. Instead, it functions as a shorthand device to represent the subjective beliefs that we currently express for the mean and variance structure over y in a way that is most convenient for the analyses that we require.

In the case study, we use quadratic surfaces of the form

$$y_i(x) = \alpha_i + \sum_j \beta_{i,j}\, x_{ij}^* + \sum_{k,l} \beta_{i,kl}\, x_{ik}^* x_{il}^* + \epsilon_i(x_i^*) + \delta_i(x) \tag{1.6}$$

where the x_{ij}^* are the components of x_i^*, though we may transform elements of x and of y before using this form. Therefore, as well as the discrepancy variances for ϵ and δ, the prior description (1.6) requires a choice of active variables, and prior means, variances and covariances over the α and β coefficients. In the case study, all linear terms $\beta_{i,j}$ in (1.6) were given positive prior variances but some quadratic terms $\beta_{i,kl}$ were eliminated from the equation.

From this prior specification, we may immediately deduce the prior mean, variance and covariance structure over $y(x)$, and it is straightforward to assess how changes in the input vector x change our beliefs over the output vector, $y(x)$. Covariances between pairs of pressure variables y_i and y_j follow from covariance specifications between the β coefficients in the two forms, and any covariance specifications between discrepancy terms. Any such covariances will usually be a decreasing function of both spatial and temporal distance between the well locations and measurement times of the two variables. Further, as we make observations on the simulator, we may routinely update our beliefs about each of the regression coefficients and the ϵ discrepancy surface, and thus about the joint beliefs over y.

The two discrepancy terms in (1.5) are as follows. First, consider our beliefs about changes in the value of a component $y_i(x)$ as we change the collection of active variables x_i^* for y_i keeping the other inputs fixed at some central values. We do not believe that the relationship between y_i and x_i^* is precisely determined by a quadratic surface, and so we add a discrepancy term $\epsilon_i(x_i^*)$ which is a function only of the active variables, which expresses our prior beliefs as to how far from the quadratic surface the value of y_i may lie. Clearly, values of $\epsilon_i(x_i^*)$ must be strongly correlated for closely neighbouring values of x_i^*. We must therefore specify a prior covariance structure over values of ϵ_i. A simple form which is sufficiently

flexible to represent all of those aspects of our prior beliefs that we wished to express is

$$\mathrm{Cov}[\epsilon_i(x_i^*), \epsilon_i(x_i^{*'})] = \sigma_{\epsilon_i}^2 \exp\left[-\theta_i(x_i^* - x_i^{*'})^T(x_i^* - x_i^{*'})\right] \tag{1.7}$$

Secondly, consider our beliefs about changes in the value of a component $y_i(x)$ as we change the whole collection of inputs x. We do not believe that such changes are entirely determined by changes in the active variables, and so we add a second discrepancy term $\delta_i(x)$, to represent our prior beliefs about variation in y_i caused by changes in the whole collection of variables. Careful specification of the covariance structure of beliefs over the values of $\delta_i(x)$ would again involve some form like (1.7) to express beliefs about local correlation. However, at this point we introduce a pragmatic simplification. Each of the other terms in (1.5) involves only the active variables for the component under consideration. It is an enormous simplification for the resulting analysis if we can preserve this property over δ_i. Therefore, when we are updating beliefs over $y_i(x)$, we will essentially treat the inputs that we have excluded from the active set as unknowns, so that we assign over $\delta_i(x)$ a constant variance $\sigma_{\delta_i}^2$ and constant covariance (in the case study this covariance was always set to zero) for any pair of different values of x. Basic to this simplification is the condition that the extra variation in $y_i(x)$ due to variation in elements of x which are excluded from x_i^* is small compared to variation in $y_i(x)$ attributable to variation in x_i^*. Provided that $\sigma_{\delta_i}^2$ is small, we lose little in our predictive description, while enormously simplifying each stage of the ensuing analysis, by this approximation.

The covariance structure for a component $y_i(x)$ is therefore determined by the choice of active variables x_i^*, functional forms $g_{ij}(x_i^*)$, the quantification of second order beliefs over the coefficients α_i and β_{ij}, the choice of constants in (1.7), and the variance of $\delta_i(x)$. The status of the various elements in (1.5) is further discussed in [CGSS96]. For our present purposes, these forms are simply intended to explain sufficient of the prior uncertainties in the problem to suggest a sensible search strategy for good history matches, while being straightforward enough for this search strategy to be tractable. At each stage, these descriptions are provisional, to be modified according to the results of diagnostic testing, and subsequently to be reassessed over sub-regions of the input space which are considered most likely to contain acceptable history matches.

7 Prior Specification

We use two sources of information for constructing prior beliefs. First, we use qualitative and quantitative judgements of reservoir engineers. Secondly, we construct and experiment on fast, approximate versions of the

simulator. These two sources are combined to quantify the prior description given in Section 6. We shall outline the basic stages of the process, describing what we now consider to be a sensible sequence of elicitation steps. This sequence was roughly followed in the case study, but, as this was very much a learning process, some of the steps were done in parallel and reconsidered several times in order to obtain a prior description.

7.1 Experts' Judgements

The first stage in the prior specification was informal discussions with a reservoir engineer from SSI who was familiar with the particular reservoir. These discussions clarified basic objectives and identified potential difficulties. Informal consideration was given to the qualitative relationships between inputs and outputs. An initial region of interest was selected as most likely to contain the reservoir geology, based around a central assessment by a geologist as to likely values for the variables.

Resulting from these discussions, we decided to vary two types of quantities to match the observed pressures. First, permeability multipliers were allowed to vary between 0.1 and 10. The variables we use to control these multipliers are on a $\log_{10}$ scale, varying from -1 to 1. Secondly, fault transmissibility multipliers may vary over a range of possible values from 0 (a sealing fault, allowing nothing to flow through) to 1 (allowing normal flow), and it was decided to use the whole range. A square root transform of the scale of the fault variable was suggested by subsequent analysis.

We also discussed the magnitudes of the measurement errors as quantified by the prior variances for y_E in (1.3). There are three ways in which pressures at wells may be measured, for each of which a prior precision was elicited, based on 90% intervals for error magnitudes which varied between 2% and 10% of observed values. For some of the wells, it was clear that the most accurate method had been used, but, in general, information about measurement procedures had not been recorded and was not easily available. For some of the readings, inferences could be drawn as to the likely method of measurement, but for most a compromise accuracy was chosen. The engineer was happy that the components of y_E should be uncorrelated with each other and with all other terms.

We further discussed the magnitude of differences between the simulator and the reservoir, as quantified by the prior variances for y_D in (1.4). It became clear that beliefs about these differences were based on rather subtle considerations. Our impression was that there was substantial prior information which could have been elicited but that this would have required considerable modelling. If our primary intention had been to make inferences about the composition of the reservoir, then this would have formed a central part of the prior structuring. However, as we were primarily concerned with the history match, we restricted attention to simple general statements of uncertainty. In particular, the median of the absolute mag-

nitudes of the percentage changes due to the components of y_D was judged to be 5% of the true values, and correlations between different components of y_D were largely judged on spatial terms with a correlation of 0.8 at a distance of 200m and negligible correlation at 1600m. However, this was complicated by faults: the effective distance between two wells separated by an active fault would be substantially greater than the physical distance, so that generally fault transmissibility should be used to modify physical distance.

The reservoir engineer further considered that the correlation structure and expected size of differences between the fast and the full simulator would be of a similar order of magnitude to those between the full simulator and the reservoir.

7.2 Using a Fast Simulator

In order to use (1.6), we need to identify a collection of active input variables for each component y_i and quantify prior beliefs about the magnitudes of the various coefficients in this prior description. This is a difficult and time consuming process, which must be repeated for each of a large number of output quantities. Further, we intend to re-examine our choices at certain stages in our search procedure. Therefore, it is very useful to have additional sources of prior information which we can use in a semi-automatic fashion to simplify the prior specification. For this reason, we constructed a fast version of the simulator, based on a coarser gridding of the reservoir, which took between ten and fifteen minutes per evaluation, as opposed to a few days. It was considered that the qualitative features of the two simulators might be sufficiently similar that collections of active variables identified on the fast simulator could serve as a choice of active variables in the prior description of beliefs for the full simulator, and that fitting models of form (1.6) on the fast simulator would give us reasonable prior means for the corresponding quantities on the full simulator.

We therefore ran the fast version of the simulator many times. In the study, we chose a 100 run Latin hypercube design (see [MCB79]) in the 40 input variables under consideration. As the study proceeded, we found that there was much additional information from the fast simulator results which could be exploited, and a larger hypercube might have been preferable. Against this, however, must be weighed the additional use of time and resources. In general, we suggest running a large hypercube at the first stage of the process and rather smaller hypercubes when we subsequently come to refit our descriptions over sub-regions.

We now have observations on 100 pairs of input and output vectors. For simplicity, we treated each component, y_i, of the output vector separately, so that for each component, we had 100 observations of the form $(x, y_i(x))$. These observations were used (i) to select a collection of active input variables for y_i, with a separate choice made for each i, (ii) to choose which

quadratic terms to include in addition to the linear terms in the prior description (1.6) and (iii) to aid in quantifying prior beliefs for use with this prior description on the full simulator.

The process for making the choice, for a given i, is as follows:

1. We first use the `stepwise` routine in S-PLUS (see [Inc93]) to fit y_i on linear and quadratic terms in all input variables by using ordinary least squares. The first six input variables to enter the description are considered as candidate active variables for y_i. This was a pragmatic choice to reduce the computational effort in the next stage of the selection process, but generally we found that there was very little residual variability left in each y_i by this stage.

2. Now we search stepwise for the 'best description' using three or fewer of the six candidate active variables. For each candidate collection of active variables, we identify stepwise the best choice of quadratic and interaction terms to supplement the linear terms in those active variables. The searches require a criterion for selection of active variables and terms. We considered various criteria for good prior descriptions in the study, seeking descriptions which are likely to be useful for restricting the range of possible inputs x for which $y_i(x)$ might be an acceptable pressure match. The quantification that was finally used was based on selecting the combination of variables which maximised, stepwise, a criterion which is based on the following intuitive motivation. Our intention in selecting active input variables is to identify, and therefore eliminate from further consideration, choices of inputs which are implausible as providing good pressure matches on the full simulator. How this will happen is that we will choose sequentially a set of observations to make on the full simulator, at each stage updating our mean and variance, for the value of $y_i(x)$ for each x, by Bayes linear fitting. As we update our beliefs, we also calculate 'implausibility' measures, which are intended to identify those collections of inputs which may be removed from the starting set R of possible input values because according to our adjusted beliefs we shall consider that it is highly 'implausible' that these values could give an adequate pressure match. Implausibility measures are discussed in Section 8 below. Focussing on component y_i, the univariate versions of the measures are based on the magnitude of the following quantity, which has the important property that it can be calculated purely on the basis of first and second-order prior beliefs assigned to the various quantities, namely

$$\frac{(\mathrm{E}_D[y_i(x) - y_{Ti}])^2}{\mathrm{Var}_D[y_i(x) - y_{Ti}]}. \tag{1.8}$$

where y_{Ti} is the ith component of y_T in (1.3), and the adjusted expectations and variances are taken using (1.1) and (1.2) based on the

data set $D = \{y_i(x_{[1]}), \ldots, y_i(x_{[n]})\}$ of values which we observe on the full simulator. We will judge that our choice has been successful if, for many values of $x \in R$, the value of (1.8) turns out to be large.

Therefore, we want to choose as active variables those for which our prior expected value for the magnitude of (1.8) is large for many values of x. To achieve this, we choose active variables to maximise our prior expectation for (1.8) as expressed by the criterion

$$\mathrm{E}\left[\int \frac{(\mathrm{E}_D[y_i(x) - y_{Ti}])^2}{\mathrm{Var}_D[y_i(x) - y_{Ti}]}\, dx\right] \tag{1.9}$$

For any fixed choice of design points, $x_{[1]}, \ldots, x_{[n]}$, and particular value x, it is possible to evaluate the prior expectation of (1.8) purely in terms of the prior covariance structures defined over $y_i(x)$ by means of (1.6). For computational purposes, the integral can then be approximated by summing this expectation over a regular grid in the active variables. Therefore, to evaluate (1.9) for a candidate collection of input variables, we must quantify prior beliefs for the various elements in (1.6) based on a combination of simple model fitting using the data that is available from our runs on the fast simulator combined with elicitation from the reservoir engineer of beliefs about the magnitude of the differences between the fast and the full simulator. This quantification is as described in step 3 below. These beliefs should incorporate any relevant data which is already available. For example, at the beginning of the study we already have the results of one run since the simulator is tested at its initial settings. Therefore, we construct beliefs which might be reasonable prior to any full simulator runs and adjust them by linear fitting on the available data to form prior beliefs for use in evaluating (1.9).

A further crucial simplification is as follows. As we will choose the input vectors at which we run the full simulator sequentially, it is a very difficult computational problem to assess our prior expectation over the actual design points that we will run on the full simulator. For this reason, we choose a simple approximation to the above criterion where we fix in advance a simple preselected design in the active variables under consideration. The idea is to choose a design which will be informative about all the coefficients in the description, without spending time choosing an optimal design for the particular coefficient estimates. For the case study, we chose designs with 11 points which depend only on the number of active input variables in the candidate description: for a single active input, the design consists of 11 equally spaced values, covering the range of the input variable; for two active inputs, we used a hypercube like design; and for three inputs, we used a design based on a cube with points at the centre

and in the middle of each face and with four other points on vertices of a smaller nested cube.

We have made a variety of somewhat ad hoc choices in order to choose the preferred collection of active variables. As yet, we have little experience as to the sensitivity of our approach to the various approximations and assumptions that we have made. However, the method appeared to select sensible collections of active variables in an automatic fashion in the present study.

3. For each component, y_i, we have now selected a collection of three active input variables x_i^*, and decided which quadratic terms in these variables to include in the relation (1.6). The quantification of prior beliefs over these quantities proceeds as follows. We begin by fitting the chosen linear description in the selected variables to the fast simulator data using ordinary least squares, assuming an uncorrelated error structure. This is a reasonable approximation, as we have a large sample on a roughly orthogonal design, with values which are reasonably well separated. The estimated coefficients $\hat{\alpha}_i$, $\hat{\beta}_{i,j}$ and $\hat{\beta}_{i,kl}$ are taken to be the corresponding prior means $\mathrm{E}[\alpha_i]$, $\mathrm{E}[\beta_{i,j}]$ and $\mathrm{E}[\beta_{i,kl}]$ for the coefficients on the full simulator, and the variance matrix of these coefficients is the sum of two components. The first is our uncertainty about the value of the estimates from the fast simulator and is taken as the standard estimate from the coefficient variance matrix. The second component represents the engineer's uncertainty about the difference between the two simulators and was taken to be diagonal with each standard deviation for the diagonal taken to be 5% of the corresponding coefficient estimate. This was subsequently increased as the case study progressed; see Section 9.

   The covariance assessments for $\epsilon_i(x_i^*)$ and $\delta_i(x)$ are estimated from the residuals from the ordinary least squares fit as follows. The overall error variance, $\sigma^2_{\epsilon_i} + \sigma^2_{\delta_i}$, is estimated by the residual variance. Individual values for $\sigma^2_{\delta_i}$ and $\sigma^2_{\epsilon_i}$ are found from examining the correlations between 'neighbouring residuals', as we always treat $\delta_i(x)$ as having zero spatial correlation. As it is hard to estimate θ_i from the data, we chose a value for θ_i which depended only on the number of active variables: 85 for one active variable, 42 for two and 32 for three active variables. This simplification is possible because, for the description building process, we linearly transform each input variable to the range $[-0.5, 0.5]$. The chosen values were arrived at by examination of the spatial autocorrelation function of the residuals from quadratic fits to cubic polynomials. Prior covariances between any of the four components α_i, β_i, ϵ_i and δ_i are all zero.

Plots of the spatial and temporal configuration of coefficients for pressures y_i with the same active variables were then used to assess how much

spatial and temporal correlation should be built into the prior specification, and to compare with the order of magnitude correlations suggested by the reservoir engineer. There is scope here for more detailed data analysis and modelling. Overall, we had considerably more confidence in our marginal specifications than in our joint specification. This, together with various computational problems, led us to prefer marginal to joint analyses in the study.

7.3 Checking the prior specification

In Sections 7.1 and 7.2, we have described the basic steps that we followed in constructing our prior description. The final stage in constructing the prior description was to check the qualitative and quantitative form of the prior specification at which we had arrived in consultation with the reservoir engineer. The form of the specification is complicated as there are many input quantities related to many output quantities. We created a graphical elicitation tool which simplified the task of the reservoir engineer, namely checking that there did not appear to be any substantial flaws in the prior beliefs that we now expressed.

Our eventual intention is that the elicitation process should become properly interactive, allowing the reservoir engineer to input a priori both qualitative belief specifications, which may be used to direct our search for active variables in the fitting stage on the fast simulator (this becomes increasingly important as the number of inputs increases), and also quantitative prior belief specifications which we may combine with the evaluations on the simulator in a consistent Bayes linear way. The elicitation tool was therefore constructed to facilitate this more general approach to elicitation. The tool that we created could be used: (i) to create and display relationships between active input variables and the corresponding outputs, basically by showing a map of the reservoir on which different inputs and outputs could be displayed and highlighted, where the elements of the display were linked to tables of variables and to time plots for the measurements; and (ii) to quantify individual aspects of beliefs, offering both graphical methods for expressing uncertainties as to the effects on a pressure output y_i from changing a geology input x_j, and also ways of grouping the variables for elicitation purposes so that we could use exchangeability arguments to extrapolate uncertainty statements from one quantity to related classes of quantities. The details of this more general methodology, and a full description as to how the graphical tool functions to support this approach, will be given elsewhere.

Only a small amount of the functionality of the graphical tool was exploited in the case study, but this was sufficient to help the reservoir engineer to grasp the qualitative structure of the prior description that we had assessed, for example by checking features such as whether similar output quantities were affected by similar active input quantities. The result of

this process was a general agreement as to the choices of active variable collections that we had made, but surprise as to how few of the fault variables had been included. This was one of the motivations which later led us to examine carefully the influence of fault variables and which led to a square root transformation of those variables; see Section 11. The graphical tool was also used to assess some simple rule of thumb prior quantifications which were roughly in accord with the prior description that we had constructed at the previous stage.

7.4 Details of the Study

The combinations of active variables chosen are shown in Table 1.2. Recall that variables 1 to 7 are permeability multipliers for the regions and that the remainder are fault transmissibility multipliers. Observe that certain inputs are active for many components; for example, x_3, the permeability multiplier for region 3, appears in all but one of the descriptions. By cross-referencing Tables 1.1 and 1.2 and Figures 1 and 2, we can observe that the choices of active variables seem to be consistent with the physical locations of wells. For example, the permeability multiplier for region 1 is x_1, which only appears for outputs at wells in that region. Similarly, outputs from wells nearer to region 2 are more likely than other outputs to have x_2 as an active variable. The fault transmissibilities which appear as active variables are usually active for outputs corresponding to wells on which the faults might be expected to have an influence.

As an example, the description developed for y_9 has two active variables, x_3 and x_4, each ranging from -1 to 1. The form of the description is

$$y_9(x) = \alpha_9 + \beta_{9,1}x_3 + \beta_{9,2}x_4 + \beta_{9,12}x_3x_4 + \epsilon_9(x_3, x_4) + \delta_9(x)$$

The estimated values of σ_{ϵ_9} and σ_{δ_9} are respectively 41 and 11. Having adjusted by the data from running the simulator at its initial settings, the means for α_9, $\beta_{9,1}$, $\beta_{9,2}$ and $\beta_{9,12}$ are respectively 2029, 213, 121 and 80. The corresponding standard deviations are 38, 13, 10 and 13. All coefficient correlations are negligible, the largest magnitude being 0.04.

Examination of the means of coefficients of x_3 and x_4 in all of these descriptions showed clear spatial and chronological structure. There was strong evidence that for wells in region 4 (the vast majority), the time of measurement could be used to help predict the coefficients. There was also clear evidence that the wells in regions 1, 2 and 3 behave differently from those in region 4, and that those in region 5 differ from both groups. On this basis, supported by the beliefs of the reservoir engineer, we decided to organise the y-variables into 12 groups, splitting them by time of measurement and by region of the reservoir in which the well is located (see column 5 of Table 1.1).

Active inputs	Outputs
3, 4, 6	8, 16, 17, 18, 20, 23, 24, 28, 30, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 72, 76
2, 3, 4	5, 55, 70, 71, 73, 74
3, 4	9, 10, 21, 22
3, 4, 31	15, 25, 27, 29
3, 4, 21	11, 12, 26
1, 2, 3	3, 54
3, 4, 7	4
3, 4, 12	14
3, 4, 17	19
3, 4, 19	1
3, 4, 26	7
3, 4, 27	13
3, 4, 35	6
3, 4, 39	31
2, 3, 7	77
3, 7, 24	2
3, 21, 25	75
4, 15, 19	32

TABLE 1.2. Collections of active input variables chosen for output variables

8 Implausibility Measures

As discussed in Section 5, we do not formulate history matching as the search for a single 'best' value of x. Instead we seek to establish how well $y(x)$ matches y_H for each x in order to discover the region R_M of acceptable matches. As complete knowledge of the surface $y(x)$ is unavailable, we build descriptions of our beliefs about the surface and, in the light of data $D = \{y(x_{[1]}), \ldots, y(x_{[n]})\}$ from runs at inputs $x_{[1]}, \ldots, x_{[n]}$ on the full simulator, update the descriptions using the Bayes linear methodology. In this section, we discuss how beliefs about the surface $y(x)$ are used to provide information about R_M. As part of our approach, we seek to eliminate from consideration values of x for which a match is implausible, so that we can obtain more accurate knowledge about y for the remaining more plausible values of x. Of course, the classification will itself be subject to uncertainty and so it is always possible that conflict between a later observation and current beliefs may force a re-classification of x-values currently identified as implausible.

We seek a summary of our beliefs about $y(x)$ which measures how implausible it is that, when evaluated, $y(x)$ will be an acceptable match. An obvious measure might seem to be $\Pr[y(x) \text{ matches } y_H \text{ acceptably} \mid D, x]$.

Implausible x would then be those for which the probability is too small. There are several obstacles to using this measure. The first is that, with the Bayes linear methodology, full probability distributions are not assessed. The second is that the definition of acceptable match is quite vague and difficult to quantify. Even were the first two difficulties to be overcome, the probability may still be misleading. For example, when $\mathrm{Var}_D[y(x)]$ is very large, the probability that $y(x)$ matches is almost certain to be small, regardless of the value of $\mathrm{E}_D[y(x)]$ or the shape of the distribution of $y(x)$. At the early stages of history matching, it is likely that $\mathrm{Var}_D[y(x)]$ will be large for almost all x. We would not wish to exclude most values of x from consideration at that stage.

The characteristics of a value of $y(x)$ which define an acceptable match depend on the purpose of the match. We have chosen to try to approximate y_T for reasons given earlier. We must also consider how to measure proximity to y_T, which should be measured differently for various components of y as quite different physical quantities are being measured and measurement methods change from time to time and may differ from one well to another. We focus on the adjusted mean and variance of $y(x) - y_T$ which we denote respectively by $\mu_D(x)$ and $\Sigma_D(x)$.

As $y(x) - y_T = y(x) - y_H + y_E$ and we consider y_E to be uncorrelated with all values of $y(x)$, then we have

$$\mu_D(x) = \mathrm{E}_D[y(x)] - y_H \tag{1.10}$$

$$\Sigma_D(x) = \mathrm{Var}_D[y(x)] + \mathrm{Var}[y_E] \tag{1.11}$$

where $\mathrm{E}_D[y(x)]$ and $\mathrm{Var}_D[y(x)]$ are the expectation and variance of $y(x)$ adjusted by D. If we had chosen to try to approximate y_H or $y(x_C)$, instead of y_T, the only changes to what follows would be in (1.10) and (1.11). Exact calculation of $\mu_D(x)$ and $\Sigma_D(x)$ requires specification of the full covariance structure between all components of y for any pair of x-values. While elicitation of meaningful coherent covariance structure is possible, it is very difficult for this problem, and the presence of highly correlated components of y is likely to make the results highly sensitive to details of the structure. There are enormous computational savings if adjustments to beliefs about a single component $y_i(x)$ use only the data from that component. This approximation has been used throughout the study.

8.1 Implausibility in One Dimension

For a single component, y_i, a simple measure of implausibility is the squared coefficient of variation

$$\gamma_i^2(x) = \mathrm{CV}_D^2[y_i(x) - y_{Ti}] = \mathrm{E}_D^2[y_i(x) - y_{Ti}]/\mathrm{Var}_D[y_i(x) - y_{Ti}] \tag{1.12}$$

which, from (1.10) and (1.11), can be calculated using only y_{Hi}, $\mathrm{Var}[y_{Ei}]$ and quantities available from Bayes linear updating of $y_i(x)$ using (1.1)

and (1.2). Large values of $\gamma_i^2(x)$ are an indication that x is implausible for any 'acceptable match' criterion which requires $y_i(x)$ to be near to y_{Ti}. Small values indicate either that $\mathrm{Var}_D[y_i(x)]$ is large or that, by Chebyshev's inequality, the probability of a match is high. As a consequence of (1.10) and (1.11), x is implausible only if the difference between $\mathrm{E}_D[y_i(x)]$ and y_{Hi} is large by comparison with the standard deviation of the error in measurement of y_{Hi}.

Using this measure, we can show that an implausible point, one with a large coefficient of variation, is expected to remain implausible. For, if D is data from the simulator (or any other collection of random quantities),

$$\mathrm{E}^2[\mathrm{CV}_D[y_i(x)]] = \mathrm{CV}^2[y_i(x)]\frac{\mathrm{Var}[y_i(x)]}{\mathrm{Var}_D[y_i(x)]} \geq \mathrm{CV}^2[y_i(x)] \tag{1.13}$$

and

$$\mathrm{CV}^2[\mathrm{CV}_D[y_i(x)]] = \mathrm{CV}^2[y_i(x)]\frac{\mathrm{Var}[y_i(x)]}{\mathrm{Var}[y_i(x)] - \mathrm{Var}_D[y_i(x)]} \geq \mathrm{CV}^2[y_i(x)] \tag{1.14}$$

where $\mathrm{CV}[\cdot]$ and $\mathrm{CV}_D[\cdot]$ denote the coefficient of variation before and after adjusting by D. The implications of (1.13) and (1.14) are that we expect the magnitude of $\gamma_i(x)$ to increase as we observe more data, and that if $\gamma_i(x)$ is currently large, it would be surprising if it were to become small in the future.

8.2 Implausibility in Many Dimensions

The way in which components of y should be combined into a single measure of implausibility is bound to be reservoir specific. For example, a contract to match a reservoir may specify that not all components need be matched, or a reservoir engineer may judge that matching some components is more important than matching others for prediction and decision making using the simulator. As a simple combined measure for multi-dimensional y, by analogy with (1.12), we define

$$\mathcal{I}(x) = \mu_D(x)^T \Sigma_D(x)^{-1} \mu_D(x) \tag{1.15}$$

where $\mu_D(x)$ and $\Sigma_D(x)$ are the mean and covariance matrix of $y(x) - y_T$ after adjustment by the data D. Large values of $\mathcal{I}(x)$ correspond to implausible values of x. The measure takes account of correlations between the components of y and also possibly different measurement scales, but does not allow for differing scales used when assessing match quality. This measure also assumes that it is important to match all components of y, as a single very poorly matching component will have a large influence.

As noted earlier, exact calculation of $\mu_D(x)$ and $\Sigma_D(x)$, and consequently of $\mathcal{I}(x)$, would require specification of the full covariance structure between

all components of y for any pair of x-values. We have chosen in this study to use approximations to $\mathcal{I}(x)$ which do not use the full covariance structure and which greatly reduce the computational burden. There are a number of reasons why computation is made quicker by not using the full correlation structure of $y(x)$. First, adjustment calculations can be carried out separately for each component of y, using only data from the same component, which immediately reduces the size of covariance matrices used by a factor of p^2 (5929 in the case study). Secondly, each component implausibility γ_i is a function of at most three components of x, which makes it possible to calculate γ_i on a grid of x-values. Thirdly, the forms of the approximations to $\mathcal{I}(x)$ allow further computational savings to be described in Section 8.5.

Let us note two extreme situations in which $\mathcal{I}(x)$ takes simple forms. If the components of y are uncorrelated, then $\mathcal{I}(x) = \sum_i \gamma_i^2(x)$ where $\gamma_i^2(x)$, the 'marginal' implausibility of x based on component y_i, is the coefficient of variation of $y_i(x) - y_{Ti}$, as in (1.12). On the other hand, if the p components of y are perfectly correlated then, using a generalised inverse of Σ_D, the implausibility is $\mathcal{I}(x) = \frac{1}{p}\sum_i \gamma_i^2(x)$. In both cases, the adjustments required to compute γ_i use only the simulator data for the i-th component.

In order to approximate $\mathcal{I}(x)$, we divide the components of y into mutually exclusive subsets, which are chosen so that the subsets are judged to be nearly uncorrelated and so that, within each subset, the components are judged to be strongly correlated. Consequently, we sum implausibilities arising from each of the subsets where the implausibility for each subset is the average of the corresponding 'marginal' implausibilities. Denoting the subsets by $S_1, S_2, \ldots$, we approximate $\mathcal{I}(x)$ by

$$\bar{\mathcal{I}}(x) = \sum_j \frac{1}{|S_j|} \sum_{i \in S_j} \gamma_i^2(x) \tag{1.16}$$

where $|S_j|$ denotes the number of components of y in S_j. This measure combines implausibilities associated with the individual components of y in a way which reflects to some extent our beliefs about their correlations.

A quite different measure also used in the case study is

$$\hat{\mathcal{I}}(x) = \max_i |\gamma_i(x)| \tag{1.17}$$

which is conservative, as it is based on the important requirement to only 'rule in' those inputs which may give an acceptable match for all components; that is, a point is implausible if it is implausible for at least one component of y. As before, if we ignore the correlation structure of y, we compute γ_i using only data values for y_i and we obtain the same computational savings as with $\bar{\mathcal{I}}(x)$. A less sensitive measure might be the m-th largest value of $|\gamma_i|$ which, used to define implausibility, would mean that implausible values of x were those for which at least m components of y

failed to match. We have not used it in this study because the computations would become rather more complex.

Whatever measure of implausibility we choose, the notion appears many times in our overall methodology: (i) to help remove from consideration implausible subsets of x values; (ii) to guide the choice of x-values for runs of the full simulator as described in Section 10, the idea being to choose as input the value of x whose y will most reduce overall uncertainty about y-values for the set of plausible x; (iii) as part of the process of choosing descriptions for the components of y, in Section 6; and (iv) to assist informally in deciding when to reformulate the description on a new sub-region of input values.

8.3 Presentation and Interpretation of Implausibility

The implausibility criteria described are functions of all of the components of x. For presentation and interpretation, we need to assess implausibility for small subsets of the components of x. It seems reasonable that a value of a single component of x is implausible only if it is implausible for all possible values of the other components. Hence, if x' and x'' are complementary subsets of the components of x, we consider, as a measure of implausibility for the components in x', the projection

$$\min_{x''} \mathcal{I}(x) \tag{1.18}$$

which is a function only of x'. We can project $\bar{\mathcal{I}}(x)$ and $\hat{\mathcal{I}}(x)$ in the same way.

Other methods for projecting implausibility measures are possible. However, in the case study, we have used only the method in (1.18), projecting onto at most three components and very often just two, to aid in the visualisation of implausibility.

As a basis for assessing implausibility based solely on our restricted belief specification we use an heuristic based on the 'three sigma rule' [Puk94], which states that for any continuous unimodal density at least 95% of the probability is within 3 standard deviations of the mean. This informally suggests that for a single component a value of $\gamma_i^2 \geq 9$ corresponds to input which is implausible as a match. As a rough calibration for $\hat{I}(x)$ in (1.16) we have informally compared values to quantiles of a χ^2 distribution having the same number of degrees of freedom as there are groups, informally treating the groups as independent. We treat $\hat{\mathcal{I}}(x)$ from (1.17) as a confirmatory check that plausibilities are similar for both plausibility methods.

There are potential dangers associated with the use of either (1.15) or (1.17). In the former, there is the danger that by combining plausibility contributions of the components of y through the generalised coefficient of variation, we may miss anomalous components or conflicts between groups of components, which also may make meaningless the overall measure of

plausibility. In the latter, there is the danger that an anomalous component will totally dominate the implausibility measure. In the case study, the individual plots of $\gamma_i^2(x)$ helped us to identify an anomalous simulated pressure and, after further examination of simulator files, it was judged reasonable to exclude it from (1.16) and (1.17). We anticipate that such aberrant behaviour may be typical in reservoir simulation studies, and suggest that it will always be worthwhile to examine, at various stages, the marginal plausibilities associated with the individual components of y. If there is good reason to distrust a component of y, it can be removed from the analysis.

8.4 Case Study Details

Figure 4 shows some of the results of our plausibility calculations for the case study, all based on adjusting the beliefs outlined in Section 7.4 by data from 8 runs of the simulator. For more information about the input settings for these runs, see Section 10.

The first four plots relate to y_9 whose only active x-variables were x_3 and x_4. The values of x_3 and x_4 for each of the 8 runs are indicated by a plus-sign on the plots. Plots (a) and (b) show respectively the adjusted expectation and adjusted standard deviation of $y_9(x) - y_{T9}$ as a function of x_3 and x_4. Plot (c) shows the adjusted coefficient of variation, $\gamma_9(x)$, of $y_9(x) - y_{T9}$ as a function of x_3 and x_4. Plot (d) shows the approximate 'match probability', to be described later in Section 10, for y_9 as a function of x_3 and x_4.

The remaining plots show the combined implausibility measures projected onto x_3 and x_4. Plot (e) shows $\bar{\mathcal{I}}(x)$, the groups used in the calculation being the ones described in Section 7.4. The contours 19, 21, 26 and 33 are the 90%, 95%, 99% and 99.9% quantiles of χ^2_{12}. Plot (f) shows $\hat{\mathcal{I}}(x)$. In both cases, we have removed y_{60} from the variables being considered. Examination of the plausibilities for individual variables suggested that y_{60} was severely out of line. Analysis of the data revealed that the values of y_{60} were extremely low and showed almost no variation. A careful examination of the simulator output files revealed that the well has basically run dry when the measurement is made. It seems that the physical behaviour of the well is poorly reproduced by the simulator.

On the basis of plot (e), as confirmed by plot (f), we decided to confine our attentions to the sub-region where $x_3 \geq -0.6$, $x_4 \leq -0.2$ and the ranges of all other components of x are unchanged. We then restarted the history matching process by running a new hypercube design on the fast simulator. For more details, see Section 11.

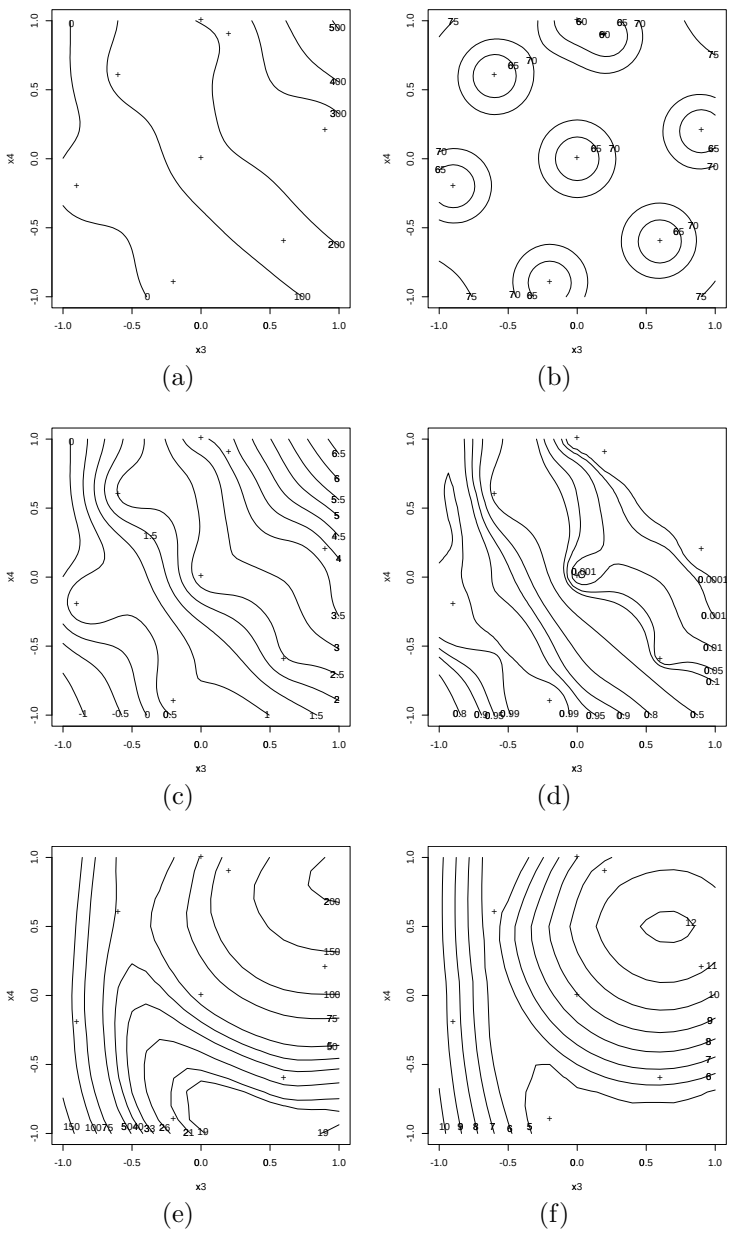


FIGURE 4. Implausibility related contour plots after 8 runs of the full simulator: (a) expectation of $y_9(x) - y_{T9}$; (b) standard deviation of $y_9(x) - y_{T9}$; (c) coefficient of variation, γ_9, of $y_9(x) - y_{T9}$; (d) 'match probability' for y_9 (see Section 10); (e) $\bar{\mathcal{I}}(x)$ projected onto x_3 and x_4; (f) $\hat{\mathcal{I}}(x)$ projected onto x_3 and x_4.

8.5 Computational Details

In this section we discuss how to compute and project our implausibility measures. Presentation in two or three dimensions requires that grids of values of a projected measure are available. A naive approach would require the calculation of grids in d dimensions followed by projection. However, d is likely to be large (40 for the case study) and the smallest possible grid requires 2^d values. A fortunate consequence of the model structure in (1.5) together with the forms of the implausibility measures (1.16) and (1.17) and the projection (1.18) is that the computations can be done easily and rapidly subject to one small assumption, namely that we never consider a value of x which is the same as any of those at which we have already evaluated y. Under this assumption, the updated values of $\gamma_i^2(x)$ depend only on the active components of x for y_i.

Since we have at most three active components of x for each y_i, it is straightforward to evaluate $\gamma_i^2(x)$ on a regular lattice of points in the active components. In practice, we have used grids having 11 evenly spaced values for each active component. We now need a way to obtain projections of the combined measures onto a grid of values in a subset of the components of x. Suppose we want to eliminate x_i from the collection:

- If x_i appears as an active variable in only a single component of y, then we simply eliminate x_i from that component of y by finding the minimum on the grid for each value of x_i. The result is a grid in the remaining active components, and we can easily eliminate all those components of x which appear just once as active components.

- If x_i appears as an active variable in more than one component of y and all share the same set of active components of x, we combine these components onto a single grid, either by taking the maximum across the components for each point on the grid or by summing the components (assuming the γ_i^2 values have already been pre-scaled to do the weighting required for the combined measure) for each point on the grid. We then replace the original group of components of y by a single virtual component of y with the new grid of values. Now x_i appears as an active variable in just one component, and we can proceed as in the previous item.

- If x_i appears as an active component in more than one component of y and those components do not all have the same active components of x, we must first put each of those components of y onto a larger grid corresponding to the joint collection of all components of x which are active for those components of y under consideration, so that they effectively all share the same collection of active components of x. We then proceed as in the previous item.

Eliminating one component of x at a time in this manner will lead to the

required projection. The only danger is that when putting a group of components of y onto a common grid in the collection of all their active components of x, we may be creating an unacceptably large grid, making computation infeasible. This situation may not always be avoidable. However, it has not been a problem for the case study and eliminating components of x in a sensible order should minimise the likelihood of trouble. We have two rules: (i) we prefer to eliminate a component which appears rarely as an active component than to eliminate one which appears frequently; and (ii) since it is simple to find what dimension of grid is required to combine a collection of components of y, we always calculate the dimension and, if it is bigger than three, we seek a different component of x to eliminate which would require a lower dimensional grid.

We may subsequently wish to find the value on the full d-dimensional grid which corresponds to a particular value in a low dimensional projection. This calculation can be done very easily by back-tracking through the projection process. All that is required is to store the intermediate grids formed when components of y are being combined during the projection process.

9 Diagnostics

As our prior description is a simplification of an extremely complex process, we may expect gross discrepancies in certain aspects of our description. Therefore, it is important to carry out diagnostic monitoring, so that we become aware of conflicts between observed simulator output and our description of beliefs about $y(x)$. A potential conflict arises when something occurs which, a priori, was considered to be very unlikely. Were we using Bayes theorem to update beliefs, the degree of conflict would be measured by calculating a suitable probability. Bayes linear analyses must use a different kind of diagnostic measure.

Instead of the probability of some chosen event, we measure the size of some chosen quantity, Z, by standardising it using its prior mean and standard deviation to obtain the diagnostic quantity $(Z - \mathrm{E}[Z])/\mathrm{Var}[Z]^{1/2}$ which we denote by $\mathrm{S}[Z]$. The prior mean and standard deviation of $\mathrm{S}[Z]$ are respectively 0 and 1. Note that the prior beliefs used to calculate $\mathrm{S}[Z]$ are not necessarily the initial beliefs but can be those from any subsequent belief adjustment prior to the observation of Z. Good candidates for diagnostic monitoring include observations and beliefs about important components used to structure beliefs.

As before, suppose that we sequentially observe the simulator outputs $y(x_{[1]}), \ldots, y(x_{[n]})$. We denote the collection of the first j observations by $D_j = \{y(x_{[1]}), \ldots, y(x_{[j]})\}$. To simplify notation, for any random quantity Z, we write $\mathrm{E}_j[Z]$ for the adjusted expectation, $\mathrm{E}_{D_j}[Z]$, of Z given ob-

servations D_j and $\mathrm{Var}_j[Z]$ for the adjusted variance $\mathrm{Var}_{D_j}[Z]$. When we evaluate $y(x_{[j]})$ on the simulator, we compare the prior expectation with the observed value. We will mainly be concerned with monitoring changes in individual components of y, and we denote the value of $y_i(x_{[j]})$ as y_{ij}. Note that, for the purposes of calculating diagnostics, we will treat $x_{[1]}, x_{[2]}, \ldots$ as known. Even if it were interesting to treat them as unknown, the properties of the sequential design process would be extremely difficult to analyse.

We judge discrepancies in an informal manner. For example, we might consider that magnitudes of $\mathrm{S}[y_{ij}]$ larger than 3 are suggestive of conflicts between our prior formulation and the data. However, the general patterns that emerge from the collection of evaluations are of principal interest, rather than any particular evaluation. For example, large values associated with all the pressures at a particular well might suggest that the fast simulator poorly approximates the behaviour of the full simulator at that well. Alternatively, occurrence of many very small values for all components of y would also be important in suggesting that we might have over-inflated the prior uncertainties, for example by over-estimating the differences between the fast and the full simulators.

We may make a more detailed assessment by monitoring the one step changes in expectation, $\mathrm{E}_{[r/]}[Z] = \mathrm{E}_r[Z] - \mathrm{E}_{r-1}[Z]$, of a quantity Z and assessing their standardised values $\mathrm{S}_{[r/]}[Z] = \mathrm{E}_{[r/]}[Z]/\mathrm{Var}_{r-1}[Z]^{1/2}$ to identify whether the sequence of changes in the predictions for Z is behaving as we expect. Of particular interest is the standardised one step comparison between an observation and its prediction given all of the preceding observations, namely

$$\mathrm{S}_{[j/]}[y_{ij}] = \frac{y_{ij} - \mathrm{E}_{j-1}[y_{ij}]}{\sqrt{\mathrm{Var}_{j-1}[y_{ij}]}} \tag{1.19}$$

The discrepancy measures $\mathrm{S}_{[r/]}[y_{ij}]$ compare changes in observation or expectation to the expected magnitudes of such changes. Thus, a change in belief, although small, may correspond to a large standardised value and be of diagnostic interest and a large change of belief may correspond to a small standardised value and have no diagnostic implication. In order to form a qualitative picture as to the nature of the changes between prior and adjusted expectations, we may decompose the collection of changes in belief as $y_{ij} - \mathrm{E}[y_{ij}] = \sum_{r=1}^{j} \mathrm{E}_{[r/]}[y_{ij}]$, where the terms $\mathrm{E}_{[r/]}[y_{ij}]$ are mutually uncorrelated so that $\mathrm{Var}[y_{ij}] = \sum_{r=1}^{j} \mathrm{Var}_{r-1}[\mathrm{E}_{[r/]}[y_{ij}]]$. Therefore, we may write $\mathrm{S}[y_{ij}]$ as

$$\mathrm{S}[y_{ij}] = \frac{\sum_{r=1}^{j} \mathrm{E}_{[r/]}[y_{ij}]}{\sqrt{\sum_{r=1}^{j} \mathrm{Var}_{r-1}[\mathrm{E}_{[r/]}[y_{ij}]]}}.$$

Thus, if $\mathrm{S}[y_{ij}]$ is large, we may identify which terms $\mathrm{E}_{[r/]}[y_{ij}]$ explain this apparent discrepancy, while if $\mathrm{S}[y_{ij}]$ is small, we can check as to whether

this is because each term $\mathrm{E}_{[r/]}[y_{ij}]$ is small, or whether some of the changes are surprisingly large and positive and some similarly negative leading to an apparently moderate standardised value.

Note that while there is a natural time ordering of the simulator runs, we may still choose to consider stepwise diagnostics based on rearrangements of the order of the runs, to see how well each observation that we have made can be predicted by all of the remaining observations.

There are various further diagnostics that we may evaluate. First, if we also evaluate the value of $y(x)$ on the fast simulator, giving a value $y'(x)$, say, then we can evaluate the standardised differences $\mathrm{S}[y_{ij} - y'_{ij}]$. Secondly, we can evaluate diagnostics on the revised assessments of the various coefficients in the description (1.5), for example evaluating

$$\mathrm{S}[\mathrm{E}_D[\beta_{ij}]] = \frac{\mathrm{E}_D[\beta_{ij}] - \mathrm{E}[\beta_{ij}]}{\sqrt{\mathrm{Var}[\mathrm{E}_D[\beta_{ij}]]}} \tag{1.20}$$

where for simplicity we compare overall changes between prior and adjusted expectations, but we might also examine stepwise changes in these values.

We have concentrated in our account on diagnostics for individual quantities as these are simple to evaluate and interpret. There is no difficulty, in principle, in evaluating joint diagnostics. For example, the analogous quantity to $\mathrm{S}[y_{ij}]$ for assessing the full observation $y(x_{[j]})$ is

$$\mathrm{S}[y(x_{[j]})] = (y(x_{[j]}) - \mathrm{E}[y(x_{[j]})])^T (\mathrm{Var}[y(x_{[j]})])^{-1} (y(x_{[j]}) - \mathrm{E}[y(x_{[j]})])$$

with a similar stepwise decomposition for changes in expectation for $y(x_{[j]})$. However, in this study, we considered that our joint prior specification for the components of $y(x)$ was not sufficiently detailed to merit careful diagnostic examination, and thus preferred to concentrate our attention on the marginal diagnostics, which are also more easily computed.

9.1 Case Study Details

Figure 5 shows boxplots of the standardised one-step ahead comparisons (1.19) for the first 8 runs of the simulator. Each boxplot corresponds to the 77 comparisons from a single run. We see that there are large standardised comparisons for all runs except the last one. The last run is extremely close in x_3 and x_4 to a preceding run and since x_3 and x_4 are very much the most important x-variables, it is hardly surprising that the diagnostics should be much smaller for what is almost a repeated measurement. Rearranging the order of runs so that in turn each run is taken as the last run does not greatly change the sizes of the standardised comparisons.

We suspected that the problem was caused by having under-estimated the difference between the full and fast simulators when building our prior description. The standard deviation of the difference between full and fast

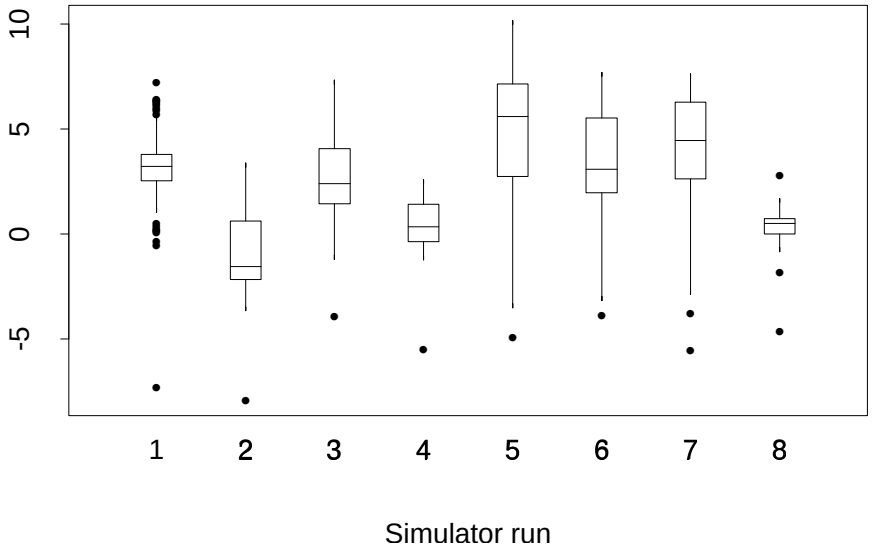


FIGURE 5. Boxplots of standardised one-step ahead comparisons for the initial 8 runs of the simulator.

simulator values for each β_{ij} had initially been taken to be 5% of the corresponding fast simulator estimate $\hat{\beta}_{ij}$. Confirmation was obtained by examining the coefficient diagnostics (1.20) which showed that the intercepts and coefficients involving x_3 and x_4 had consistently large diagnostics. We ignored other coefficients as little information was available from the simulator runs. We decided that for subsequent calculations, in place of the 5% figure, we would use a value of 20%. The 5% figure supplied by the engineer was a median absolute deviation for the difference between the simulators of each component of $y_i(x)$. For a normal distribution, this corresponds to a standard deviation of approximately 7%. In addition, each description has several coefficients (typically nine) and, if differences between full and fast simulator coefficients are independent with a standard deviation of 20% and have equal magnitudes of effect on y_i, then the standard deviation of the change in $y_i(x)$ would be approximately 20%. This value was used when building the new description from the fast simulator hypercube in the sub-region and in subsequent design calculations.

10 Choosing Runs

We used three methods of choosing inputs for runs on the full simulator in the course of the study.

x_3	0.0	0.9	0.6	0.2	−0.2	−0.6	−0.9	0.0
x_4	0.0	0.2	−0.6	0.9	−0.9	0.6	−0.2	1.0
x_6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0

TABLE 1.3. Input values for x_3, x_4 and x_6 for the initial 8 simulator runs

Stage 1

The output of a simulator run at initial settings was available; this was our first run. We then chose a simple design varying a small number of input variables which were thought to be important in determining many of the output variables. In particular, to save time, the next six runs were chosen as an orthogonal design in the two permeability components x_3 and x_4 (which featured in most of the descriptions for the 77 pressures), with the other components fixed at their initial values. Informal inspection of marginal implausibility plots suggested that a further run, varying x_6 as well as x_3 and x_4 would be useful. The input settings for x_3, x_4 and x_6, used for these first 8 runs, are shown in Table 1.3.

Stage 2

Our second method for the subsequent runs was a sequential search, maximising a criterion based on the approach described in [CGSS96]. The idea is to choose the next run of the full simulator to be at that input value which most reduces uncertainty about outputs at those input values currently not judged by our implausibility criterion to be implausible. Formally, we find the value of $x_{[j+1]}$ for the next design point which maximises

$$\sum_i l_i \int \mathrm{Corr}^2_{[j]}[y_i(x_{[j+1]}), (y_i(x) - y_{T_i})] w_i(x)\, dx \tag{1.21}$$

where $\mathrm{Corr}_{[j]}[\cdot,\cdot]$ denotes the correlation adjusted by values of $\{y_{i1}, \ldots, y_{ij}\}$ of the pressure component y_i from the previous j runs of the full simulator.

The anatomy of the design criterion in (1.21) is as follows. First, for a given input vector x, the correlation in the integrand is a measure of how much we learn about the component $y_i(x) - y_{T_i}$ when we observe $y_i(x_{[j+1]})$ — the larger the correlation, the more we learn. Secondly, as we are not interested in learning about implausible x, we discount such inputs by means of a weight function $w_i(x)$ for y_i. In the case study, the weight function was taken to be the simple bounded function $w_i(x) = \exp(-\frac{1}{2}\gamma_i^2(x))$, which gives low weight to implausible x, as measured by the implausibility measure $\gamma_i^2(x)$ given in (1.12). Other choices of weight function $w_i(x)$ are available depending on the extent to which we wish to discount relatively implausible points in our criterion; see, for example, the choice in [CGSS96]. Thirdly, the integral over x is then an overall measure of

how much we can learn about the component y_i when observing $y_i(x_{[j+1]})$. Fourthly, the final criterion in (1.21) is formed as a weighted sum of the component measures, the weights l_i reflecting the relative importance of the individual components of y. For simplicity, in the case study, the weights were taken to be the same as those used for $\bar{\mathcal{I}}(x)$ in (1.16), so that l_i is the reciprocal of the number of components of y in the group containing y_i.

When calculating the criterion, we assume that $x_{[j+1]}$ differs from each of $x_{[1]}, \ldots, x_{[j]}$ and also that the points we wish to learn about all differ from $x_{[j+1]}$ in some component. Then the integrand in (1.21) is affected only by the active components of x for y_i so we can integrate over all other components, thereby introducing a multiplicative constant. The integral in (1.21) cannot be evaluated exactly, either in closed form or computationally. Instead, we approximate it using the average over a lattice of values of the active components of x. We have used lattices with 11 equally-spaced values per component.

The criterion is readily evaluated for any choice of $x_{[j+1]}$. However, the problems involved in optimisation of a continuous objective function of many variables are intimidating given that we know little about the geometry of the objective function of the design criterion. Instead we attempt to find the optimum on a small rectangular lattice of values.

We proceed in two stages. Denote by l_i and r_i the lower and upper bounds for x_i. Then we search, first for the optimum over the lattice where each x_i takes the values $l_i + (r_i - l_i)/6, l_i + (r_i - l_i)/2, l_i + 5(r_i - l_1)/6$, and secondly, over the lattice where each x_i takes the values $x_i^{(0)} - (r_i - l_i)/6$, $x_i^{(0)}$ and $x_i^{(0)} + (r_i - l_i)/6$, where $x_i^{(0)}$ is the optimum value for x_i from the first stage. This is an approximation to finding the best point on a lattice where each x_i takes seven values.

However, even with such small lattices, the direct search for the optimum may require up to 3^d evaluations of the criterion; for example in the case study there were eighteen active inputs for the initial region. Instead we optimise by alternating between two groups of components of x (in the case study, one with ten components and the other with eight) until the optimal point does not change. We start by optimising with respect to the most common active variables with the others being left set at their original simulator settings, if that is possible, otherwise at their central values. Each stage of the alternating procedure requires us to perform a relatively easy optimisation on a smaller lattice.

Stage 3

For the final run in the case study, we sought an indication of the overall degree of success of the matching procedure. We decided to choose our input to maximise the expected number of matching components. An engineer, when shown $y(x)$, might consider the i-th component to match if $|y_i(x) - y_{Hi}| \leq C_i \sigma_{y_{E_i}}$; that is, if $y_i(x)$ is within C_i measurement error

standard deviations of the corresponding historical value. The choice of C_i might be made according to an engineering criterion of match quality or some standard value might be used. As an approximation, we calculate the probability of this event assuming that $y_i(x)$ has some convenient distribution parametrised by its mean and standard deviation. The sum of these probabilities is the expected number of matching components of y.

We can project this expectation into a smaller number of components of x in the same way as we did for implausibility but this time taking the maximum over the components being eliminated. To choose the point to evaluate, we find the value of x which has the highest expected number of matching components. This calculation can be done by back-tracking through the projection process having projected onto a couple of variables and having found the best value for those variables. All that is required is to store the grids formed by combining variables at each stage of the projection.

11 Case Study: Further Details

We have described how our prior descriptions were formed, how they were used to select runs on the simulator, how we used plausibility maps to restrict the original region to a smaller sub-region, and how diagnostics caused us to change our prior description. We now give further details of the application of the methodology to the sub-region where $x_3 \geq -0.6$, $x_4 \leq -0.2$ and where all other components of x vary over the original range of $[-1, 1]$.

We ran a fast simulator hypercube of size 100 on the sub-region and built descriptions using the process outlined in Section 7.2. In stage 2 of this process, we took the two previous runs which fell in our sub-region as available data, and in stage 3 took the standard deviation of the difference in coefficient values for the simulators to be 20% of the magnitude of the coefficient estimate from the fast simulator data instead of the original 5%.

Data analysis of the least squares fits suggested a square root transformation for x components representing transmissibilities of faults, as this results in a better quadratic fit. Descriptions were then rebuilt using the same fast simulator data. It should be noted that, as a result, the hypercube is no longer orthogonal in the new coordinate system.

The collections of active variables chosen for components for the sub-region differed substantially from those for the original region, but in a manner consistent with the change of region. For many components, x_3 or x_4 was replaced by a permeability multiplier for a nearby region or a transmissibility multiplier for a nearby fault. Where the same active variables were chosen, their order of importance often changed. Means of coefficients for terms also changed substantially, but usually without changing sign.

	Simulator runs									
	3	5	9	10	11	12	13	14	15	Trial
x_1	0.00	0.00	−0.67	−1.00	0.33	−0.67	−0.33	1.00	1.00	−0.80
x_2	0.00	0.00	−0.67	1.00	−0.67	1.00	−1.00	1.00	−0.67	0.00
x_3	0.60	−0.20	0.20	1.00	−0.60	−0.60	0.73	−0.33	0.20	−0.44
x_4	−0.60	−0.90	−0.60	−1.00	−0.20	−0.87	−0.20	−0.20	−1.00	−0.92
x_5	0.00	0.00	0.67	−0.33	−1.00	−0.67	1.00	0.67	0.00	0.00
x_6	0.00	0.00	0.33	−0.67	−0.67	1.00	−1.00	1.00	−1.00	0.40
x_7	0.00	0.00	1.00	0.33	−1.00	1.00	−1.00	1.00	1.00	1.00
x_8	1.00	1.00	0.44	0.44	0.00	0.25	0.03	0.00	0.69	0.04
x_{10}	1.00	1.00	0.11	0.11	0.25	0.00	0.03	0.44	0.00	0.00
x_{19}	1.00	1.00	0.00	0.03	0.25	0.44	1.00	0.00	0.00	1.00
x_{23}	1.00	1.00	0.00	0.11	0.25	0.11	0.25	0.11	0.11	0.49
x_{25}	1.00	1.00	0.00	0.11	0.25	0.00	0.00	1.00	0.00	1.00
x_{31}	1.00	1.00	0.00	0.11	0.03	0.03	0.00	1.00	0.00	0.00
x_{33}	1.00	1.00	0.44	0.25	0.03	0.00	1.00	0.00	0.69	0.49
x_{35}	1.00	1.00	0.25	0.44	0.03	0.00	1.00	0.00	0.00	0.00
x_{36}	1.00	1.00	0.00	0.03	0.25	0.00	0.03	0.69	0.44	0.00
x_{37}	1.00	1.00	0.03	0.03	0.11	0.00	0.00	1.00	0.69	0.00
x_{39}	1.00	1.00	0.03	0.00	0.25	0.03	0.00	0.00	0.00	0.00

TABLE 1.4. Input values for the simulator runs in the sub-region

We then chose runs on the full simulator sequentially according to the design criterion described in Section 10. After a total of 9 runs shown in Table 1.4, 2 from the original region (runs 3 and 5 in Table 1.3) and 7 from the sub-region chosen according to our design criterion, we felt that our plausibility measures had stabilised. While it is difficult to provide a detailed qualitative interpretation of the choices made for the 7 new runs, some pattern does emerge. The values of x_3 and x_4 are shown superimposed on both panels in Figure 6, and cover most of the sub-region for those variables. The other permeability multipliers all vary across their entire range, taking some non-extreme values. About half of the fault transmissibility multipliers vary across their entire range, the others being selected in sealing or partially open configurations. However, the latter faults were fully open in the 2 runs inherited from the original design and so much of the input space has been explored.

Diagnostic analysis along the same lines as in Section 9.1 showed generally large diagnostics for observations and coefficient means, suggesting that the figure of 20% of coefficient mean, used as the standard deviation for differences in coefficients between the fast and full simulators, was still too small. We tried increasing it from 20% to 50% which stabilised the diagnostics. Subsequent calculations and figures presented hereafter were

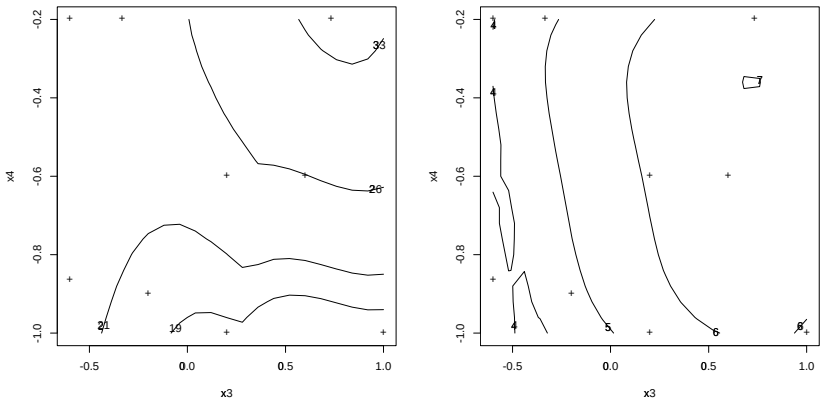


FIGURE 6. Contour plots of combined plausibility for the sub-region, after 15 simulator runs and excluding y_{60}, projected onto x_3 and x_4: (a) $\bar{\mathcal{I}}(x)$, (b) $\hat{\mathcal{I}}(x)$.

based on this value.

Figure 6 shows the new combined plausibility plots. Panel (a) shows $\bar{\mathcal{I}}(x)$ using the same contours as in Figure 4(e). We see that by this measure, large parts of the sub-region remain plausible. However, panel (b) shows $\hat{\mathcal{I}}(x)$ which is quite different in character. The differences in shape between the two panels caused us to look again at marginal implausibilities, which suggested that, in particular, components 46 and 77 of y were anomalous. Were we to judge by the marginal plausibility plots, we would either regard the entire region as implausible, or decide that the observed pressures at those wells were suspect, or decide that the simulator was unable to reproduce those pressures.

At this point, in order to judge our methodology by the standards of conventional history matching, we decided to choose a potential match point in the sub-region to see how many matches we might get. We used the method described in stage 3 of Section 10, taking $C_i = 2$ for all components and using a Gaussian distribution to approximate the required probabilities. The simulator inputs chosen are shown as the last column of Table 1.4.

Summaries of the results of all 16 simulator runs are shown in Figure 7. Each boxplot represents 76 values of $(y_i - y_{Hi})/\sqrt{\mathrm{Var}[y_{Ei}]}$, y_{60} having been omitted. The top panel uses the original measurement error standard deviations supplied by the engineer. However, many of those standard deviations were chosen as a compromise between two values as the method of measurement was uncertain. The bottom panel shows the consequences of replacing all those standard deviations by the larger of the two possible values.

Apart from a small number of wells, the match achieved on the 16th run appears generally satisfactory. The engineer at SSI considered the overall

match quality to be acceptable according to current practice. It is a fairly common experience in history matching that there are a small number of wells for which special modifications to the simulation model are required. There is some evidence that the problems relate to wells perforated only in the bottom layer or located close to the edge of the simulation region.

12 Concluding Comments

We have further developed the methodology described in [CGSS96] and applied it to a case study in which the aim has been to match historical well pressures from a large gas producing reservoir. The reservoir has previously been studied by our industrial partner SSI as part of a consultancy project for the companies who operate the reservoir.

As is common in real applications, there have been number of unforeseen difficulties and false starts, some of them methodological but most of them technical, which have imposed serious time constraints. For example, as we were developing and applying the methodology simultaneously, some of the steps in our procedure were taken out of our prescribed order. However, we have successfully implemented key features of our overall methodology and are confident that we have a useful approach to a class of inverse problems of which 'history matching' is an example; namely, those involving computer experimentation to tune CPU-expensive code for complex models of natural phenomena to noisy field data.

Two important issues arising from this study are: (i) the need to distinguish carefully between eliminating implausible inputs and selecting a matching input, for which the criteria will be different; and (ii) the importance of assessment of measurement error associated with individual pressures.

Some of the choices made in our approach are based on somewhat ad hoc criteria, which could doubtless be made more efficient. Additional features of the problem which could benefit from further study are as follows.

(i) The quality of the prior information elicited from reservoir engineers.

(ii) Learning about the difference between the two simulators based on the observed differences between the fast and full simulators on individual runs.

(iii) General aspects of combined plausibility measurement and display.

(iv) Appropriate levels of complexity for the fitted descriptions. For example, we might start by fitting simpler descriptions.

(v) The difference between the reservoir and the full simulator, which we have mostly ignored in this account. We plan to consider this more fully elsewhere.

(vi) Match quality. For example, we might monitor variables which are not formally included in the matching process or examine trends.

(vii) Use of detailed prior information about reservoir properties.

(viii) More decision theoretic views as to how many runs we should take, balancing costs and benefits of continuing or stopping by quantifying the value of the match for prediction using the simulator and for learning about the reservoir.

(ix) The integration of the various stages of our methodology into a single tool. This would seem desirable if the approach is to be used for decision support.

(x) The advantages and disadvantages of using joint versus marginal covariance structures. In particular, it would be useful to develop a coherent covariance structure for the different pressures that properly reflects their spatial and temporal features.

(xi) The dramatic variation in simulator run time for different inputs. Our decision choices should reflect this variation.

Acknowledgements

We are especially grateful to Adam Little and Bob Parish at Scientific Software-Intercomp (UK), our industrial collaborators, for their support and for making available the case study and reservoir simulator. This work has been supported by a United Kingdom Engineering and Physical Sciences Research Council grant under their Complex Stochastic Systems Initiative. We would like to thank the referees and editor for their careful reading of the paper.

13 References

[CGSS96] P.S. Craig, M. Goldstein, A.H. Seheult, and J.A. Smith. Bayes linear strategies for matching hydrocarbon reservoir history. In Berger et al., editors, *Bayesian Statistics 5*. Oxford, 1996.

[FG93] M. Farrow and M. Goldstein. Bayes linear methods for grouped multivariate repeated measurement studies with application to crossover trials. *Biometrika*, 80:39–59, 1993.

[Gol96] M. Goldstein. Prior inferences for posterior judgements. In M. L. D. Chiara et al., editors, *10th International Congress of Logic, Methodology and Philosophy of Science*. Kluwer, to appear, 1996.

[Inc93] Statistical Sciences Inc. *S-PLUS Reference Manual, Version 3.2*. StatSci, a division of MathSoft Inc., Seattle, 1993.

[MCB79] M.D. McKay, W.J. Conover, and R.J. Beckham. A comparison of three methods for selecting values of input variables in the analysis of output of computer code. *Technometrics*, 21:239–245, 1979.

[MD90] C. C. Mattax and R. L. Dalton. *Reservoir Simulation*. Society of Petroleum Engineers, 1990.

[Puk94] F. Pukelsheim. The three sigma rule. *The American Statistician*, 48:88–91, 1994.

[SWMW89] J. Sacks, W. Welch, T. Mitchell, and H. Wynn. Design and analysis of computer experiments. *Statistical Science*, 4:409–435, 1989.

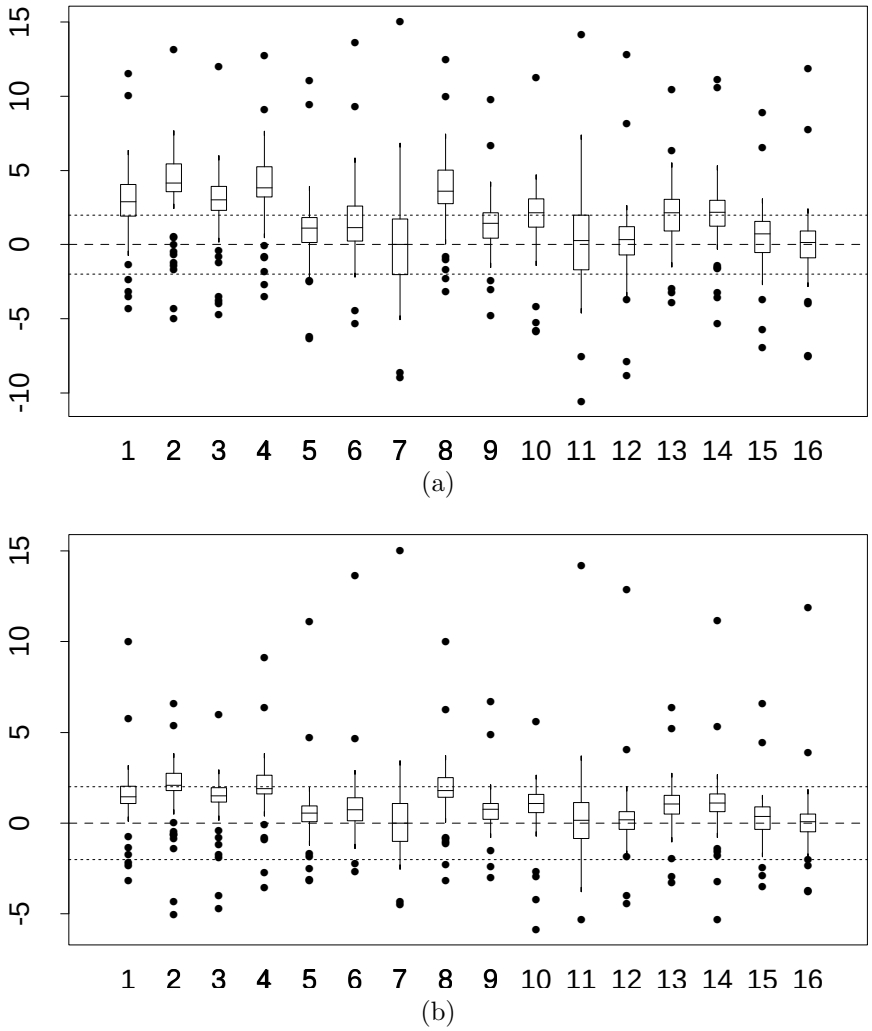


FIGURE 7. Boxplots, for all 16 runs, of deviations of simulator output from history, scaled by measurement error standard deviation: (a) using the engineer's compromise standard deviation for those pressures for which the measurement method is uncertain; and (b) using the larger of the two possible standard deviations for the same pressures.