
Article:
Owen, J. orcid.org/0009-0000-9618-9589 and Vernon, I. (2025) Bayesian emulation of grey-box multimodel ensembles exploiting known interior structure. SIAM/ASA Journal on Uncertainty Quantification, 13 (3). pp. 1501-1542. ISSN: 2166-2525

https://doi.org/10.1137/24m1669037

Bayesian Emulation of Grey-Box Multi-Model Ensembles Exploiting Known Interior Structure*

Jonathan Owen† and Ian Vernon‡

Abstract. Computer models are widely used to study complex real world physical systems. However, there are major limitations to their direct use including: their complex structure; large numbers of inputs and outputs; and long evaluation times. Bayesian emulators are an effective means of addressing these challenges providing fast and efficient statistical approximation for computer model outputs. It is commonly assumed that computer models behave like a "black-box" function with no knowledge of the output prior to its evaluation. This ensures that emulators are generalisable but potentially limits their accuracy compared with exploiting such knowledge of constrained or structured output behaviour. We assume a "grey-box" computer model and develop a methodological toolkit for its analysis. This includes: multi-model ensemble subsampling to identifying a representative model subset to reduce computational expense; constructing a targeted Bayesian design for optimisation or decision support; a "divide-and-conquer" approach to emulating sums of outputs; structured emulators exploiting known constrained and structured behaviour of constituent outputs through splitting the parameter space and imposing truncations; emulation of sums of time series outputs; and emulation of multi-model ensemble outputs. Combining these methods establishes a hierarchical emulation framework which achieves greater physical interpretability and more accurate emulator predictions. This research is motivated by and applied to the commercially important TNO OLYMPUS Well Control Optimisation Challenge from the petroleum industry which we re-express as a decision support under uncertainty problem. We thus encourage users to examine their "black-box" simulators to achieve superior emulator accuracy.

Key words. Computer models, Bayesian emulation, Bayes linear, Known simulator behaviour, Multi-model ensembles, Decision support under uncertainty

MSC codes. 62J, 62K, 62P

1. Introduction.

Mathematical models of complex real world physical systems in the form of numerical codes known as computer models or simulators are prevalent across many scientific disciplines, industry, and government. They are used to: study the dynamics of physical systems [59]; calibrate or history match to observation data [9, 10]; and to guide decision making processes [42, 32]. However, computer models commonly exhibit a complex structure; possess large numbers of inputs and outputs, including spatial-temporal fields; and crucially have a high computational expense of evaluation. In order to address such challenges, a suite of Bayesian uncertainty analysis methodology has been developed for using computer models to perform inferences about real world systems. Of principal importance are Bayesian emulators, also known as surrogate models, which provide fast statistical approximations to (functions of) the computer model outputs for as yet unevaluated parameter settings, along with a corresponding statement of the associated uncertainty [9, 37, 57]. They are typically many orders of magnitude quicker to evaluate than the computer model. Emulators have

*Resubmitted to the editors 30th March 2025.

†School of Mathematical and Physical Sciences, University of Sheffield, United Kingdom (jonathan.owen@sheffield.ac.uk).

‡Department of Mathematical Sciences, Durham University, United Kingdom (i.r.vernon@durham.ac.uk).

been successfully employed across a wide range of applications including: climate science [21, 49, 63, 14]; cosmology [57, 56, 27, 31]; epidemiology [3, 4, 60]; and petroleum reservoir engineering [10, 11, 12, 9, 38, 13].

Emulation is frequently based on the assumption that a computer model behaves like a "black-box" function: the output at a given parameter setting is unknown prior to model evaluation; as well as users' possessing no insight of the structure or links between individual physical processes. Whilst this assumption ensures that emulation methodology is generalisable, it potentially limits the emulator accuracy compared to when a user has an understanding of how certain outputs behave with respect to changes in the inputs. In this paper we assume a "grey-box" simulator which we define as possessing insight into the model behaviour prior to its evaluation, but without any specific knowledge of the underlying physics, model structure or equations governing the model. Physics informed approaches encompass: emulators for functions with known boundaries [58, 29]; emulators for functions possessing structured (partial) discontinuities in their input parameter space [41, 61]; and physics-informed neural networks which encode prior information of physical laws in non-linear partial differential equations which are used in the loss function when fitting a neural network [45], however these are unsuitable for application to "grey-box" computer models of the described form.

In this paper we address the problem of constructing an accurate and efficient emulator for an output of interest obtained from a multi-model ensemble whilst exploiting known behaviour of individual ensemble members and components of the output. Combining these methods yields a novel hierarchical emulation framework and toolkit to incorporate specific features in the analysis of "grey-box" computer models resulting in a physically interpretable emulator with accurate predictions. The toolkit includes:

1. Subsampling from multi-model ensembles technique to identify a representative subset of models enabling more efficient use of available computational resources (section 4),
2. Targeted Bayesian design of computer model simulations geared towards the optimisation or decision support objective (section 5),
3. Divide-and-conquer approach to emulation where the output of interest is represented by a linear combination of constituent model outputs (section 6),
4. Structured emulation of "grey-box" computer models exploiting a known form of simulator behaviour to divide the parameter space by mode of behaviour and employing emulator truncation to enforce known physical effects (section 7),
5. Emulation of outputs formed as sums of time series outputs (section 8),
6. Emulation of ensemble mean outputs (section 9).

All of these methods were required to address the challenges encountered in the motivating problem, although for the analysis of other computationally expensive computer models and multi-model ensembles it may be appropriate to employ a subset of these techniques.

This research is motivated by the highly complex and commercially significant TNO OLYMPUS Well Control Optimisation Challenge [54, 55] from the petroleum industry. The aim is to maximise the expected Net Present Value (NPV) objective function over the field lifetime with respect to well control decision parameters (target production and injection rates), whilst accounting for geological uncertainty represented through a multi-model ensemble of 50 realisations from an underlying stochastic geology model. We recast this as a decision support problem for which emulation and Bayesian uncertainty analysis techniques are essential due to

the computational expense of the ensemble and high-dimensionality of the decision parameter space. An initial attempt at the TNO OLYMPUS Well Control Optimisation Challenge from a Bayesian statistical perspective is presented in [42] which was only moderately successful as it failed to exploit a number of challenging features in the multi-model ensemble mean NPV output. We therefore propose a series of methodological advances in this paper for which all were required in the TNO OLYMPUS Well Control Optimisation Challenge. Implementation of the efficient multi-model ensemble subsampling technique to identify a representative subset of models constitutes a novel application to petroleum reservoir engineering and greatly reduces the computational expense of the analysis. For each ensemble member the NPV is computed as the sum of discounted time series model outputs. Many of these exhibit known constrained or structured behaviour with respect to their corresponding well control parameters for which we formulate structured emulators. These are combined within our hierarchical emulator construction. We demonstrate a notable reduction in the emulator uncertainty compared with uninformed Bayes linear emulators. Whilst we establish our techniques in the context of decision support for well control optimisation under uncertainty, the overarching framework is flexible and adaptable to handle other structured forms of simulator outputs.

In section 2 we describe the motivating TNO OLYMPUS Field Development Optimisation Challenge. Section 3 provides an overview of Bayesian emulation methodology and an application of Bayes linear emulators to the TNO OLYMPUS Well Control Optimisation Challenge ensemble mean NPV objective function. For sections 4 to 9, as in the above numbered list, we detail each methodological developments and its application to the TNO OLYMPUS Well Control Optimisation Challenge in turn. In section 9 a comparison is also performed of the developed approach to emulation with direct Bayes linear emulation of the ensemble mean NPV. A conclusion and future research directions are discussed in section 10.

2. TNO OLYMPUS Well Control Optimisation Challenge.

First we provide an overview of the TNO OLYMPUS Well Control Optimisation Challenge in subsection 2.1 before performing an exploratory analysis to highlight several important features of this challenge in subsection 2.2 which motivate the subsequent methodological development.

2.1. Summary.

A major and commercially important challenge in the petroleum industry is field development under uncertainty for a green oil field[1]. The Netherlands Organisation for Applied Scientific Research (TNO), as part of Integrated Systems Approach for Petroleum Production (ISAPP) research programme, devised the TNO OLYMPUS Field Development Optimisation Challenge [55] (abbreviated to TNO OLYMPUS Challenge) to encourage research and technological advancements to address the problem of optimisation under uncertainty. There is a particular emphasis on the uncertainty induced by the unknown underlying geology. The TNO OLYMPUS Challenge has received much attention across academia and industry with results of the competition phase presented at the EAGE/TNO Workshop on OLYMPUS Field Development Optimization [15].

The TNO OLYMPUS Challenge is based around the fictitious oil reservoir named OLYMPUS (inspired by a virgin oil field in the North Sea) and specifically designed by TNO for the

[1]A green oil field is a new subsurface region believed to contain oil or gas which has yet to be exploited meaning that no drilling, production or injection has been performed.

challenge. OLYMPUS is a medium complexity model of size 9km by 3km, with a depth of 50m split into 16 layers for modelling purposes. The design was conceived to imitate a real oil field possessing many of the features encountered in actual oil fields including: boundary and minor geological faults; two vertical zones separated by an impermeable shale layer (the top layer contains fluvial channel sands embedded in floodplain shale, whilst the bottom layer consists of alternating layers of coarse, medium and fine sands); as well as multiple types of facies (body of rock of specified characteristics) including channel sands, shale, and multiple types of sand. Geological uncertainty (unknown porosity, permeability, net-to-gross, and initial water saturation) is represented via a multi-model ensemble of $N = 50$ OLYMPUS realisations of a stochastic geology model. These are labelled as OLYMPUS 1 to 50. Full details of the model can be found in [54].

The TNO OLYMPUS Challenge consists of three sub-challenges:

1. Well control optimisation,
2. Field development optimisation,
3. Joint optimisation of well placement and well control.

In this paper we focus on the first where the aim is to develop an optimal strategy with respect to maximising the expected Net Present Value (NPV) objective function over the 20 year field lifetime (starting January 1, 2016) with accumulation and discounting at 3 month intervals. The NPV for an individual OLYMPUS model is denoted $\text{NPV}_j(\mathbf{d})$ and is defined in (2.1) as a function of a vector of decision parameters, $\mathbf{d}$, consisting of target production and injection rates for producer and injector wells respectively.

$$\text{NPV}_j(\mathbf{d}) = \sum_{i=1}^{N_t} \frac{R_j(\mathbf{d}, t_i)}{(1+d)^{\frac{t_i}{\tau}}} \tag{2.1}$$

$$\mathbb{E}[\text{NPV}](\mathbf{d}) \approx \overline{\text{NPV}}(\mathbf{d}) = \frac{1}{N} \sum_{j=1}^{N} \text{NPV}_j(\mathbf{d}) \tag{2.2}$$

The index i refers to the time interval $\Delta t_i = t_i - t_{i-1}$, total number of time intervals N_t, fixed discount factor $d = 0.08$, time interval for discounting $\tau = 365$ days, and $R_j(\mathbf{d}, t_i)$ as the difference of all revenue and expenditure during the interval Δt_i, and $j = 1, \ldots, N = 50$ indexes the particular OLYMPUS realisation of the stochastic geology model. The expected NPV is approximated by the ensemble mean NPV defined in (2.2), as dictated by the TNO OLYMPUS Challenge, and hence forms our quantity of interest.

For the well control optimisation challenge a fixed well configuration is provided by TNO based on oil reservoir engineering principles with $R_j(\mathbf{d}, t_i)$ defined in (2.3), where $Q_{j,op}(\mathbf{d}, t_i)$, $Q_{j,wp}(\mathbf{d}, t_i)$, and $Q_{j,wi}(\mathbf{d}, t_i)$ are the Field Oil Production Total (FOPT), Field Water Production Total (FWPT), and Field Water Injection Total (FWIT) volumes in time interval Δt_i under controls $\mathbf{d}$ respectively.

$$R_j(\mathbf{d}, t_i) = Q_{j,op}(\mathbf{d}, t_i) \cdot r_{op} - Q_{j,wp}(\mathbf{d}, t_i) \cdot r_{wp} - Q_{j,wi}(\mathbf{d}, t_i) \cdot r_{wi} \tag{2.3}$$

The analogous quantities for an individual well are labelled as WOPT, WWPT, and WWIT respectively. TNO stipulate fixed oil revenue $r_{op} = 45$ \$ per bbl (where bbl are the units for

a standard barrel of oil, approximately 159L), water production cost $r_{wp} = 6$ \$ per bbl, and water injection cost $r_{wi} = 2$ \$ per bbl.

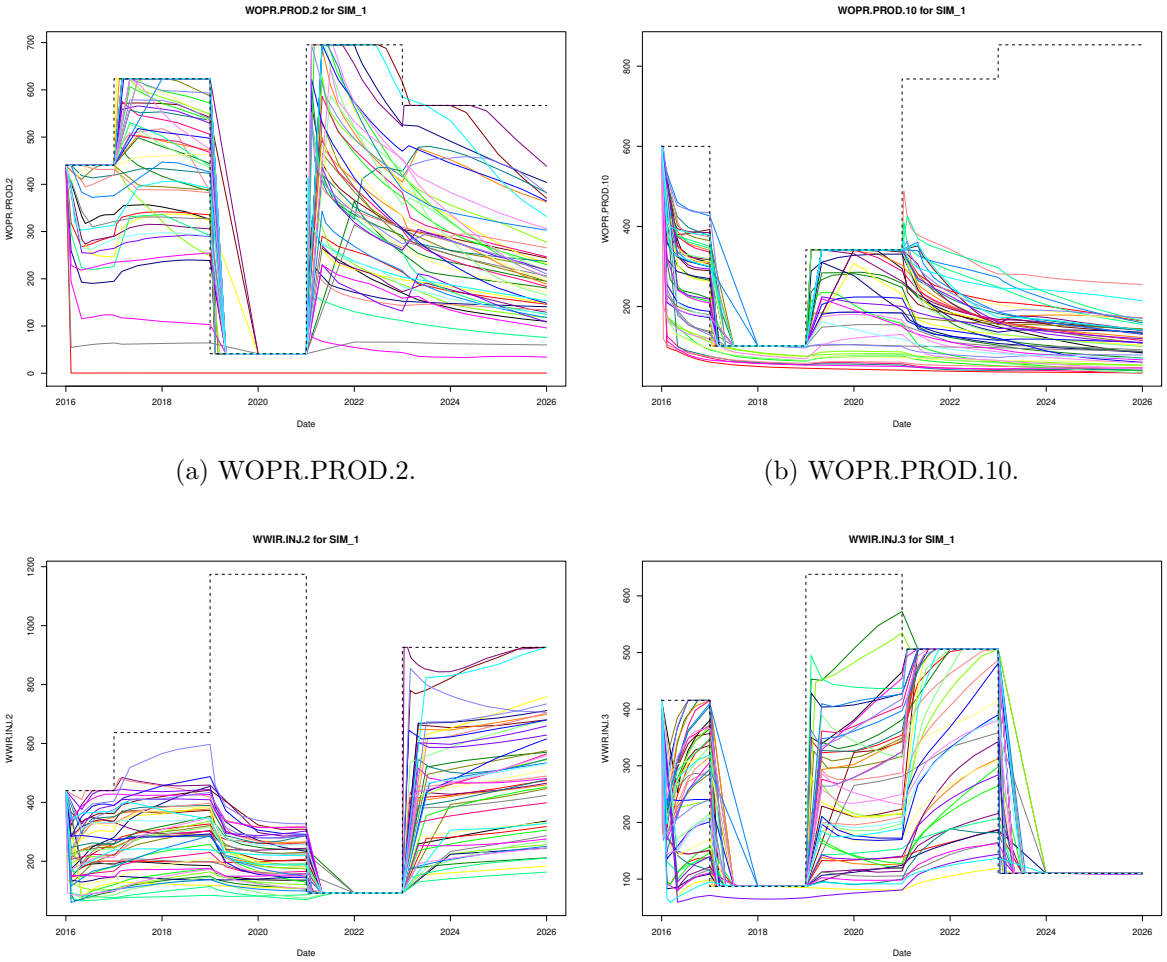
For demonstrative purposes we focus on the control of a subset of the wells enclosed between two partial fault boundaries and in close proximity consisting of two producer wells: 2 & 10, and two injector wells 2 & 3, with eight control intervals starting on January 1, 2016, 2018, 2020, 2022, 2024, 2026, 2028 & 2032; thus a total of $D = 32$ decision parameters. Throughout we represent specific individual decision parameters by d_{jk,t_i}, where $j \in \{P, I\}$ refers to the well type (P producer, I injector), k is the well number, and t_i is the control interval start date. Note that each control interval consists of multiple 3 month discounting periods. Collectively these wells are referred to as the Controlled Wells Group (CWG) which provides a sub-problem of interacting wells on which to illustrate the presented methodology. All remaining wells within OLYMPUS use the fixed controls specified in the TNO reference strategy [55]. The expected NPV objective function is computed from contributions of wells in the CWG only.

We believe that the TNO OLYMPUS Challenge setup does not faithfully represent the real world field development under uncertainty problem where ensembles of computer models are used to aid decision makers. Instead, there is an emphasis on developing efficient ensemble optimisation algorithms to identify a single optimal strategy. A full critique and discussion of these limitations is presented in [40, Sec. 3.1] and [42]. We therefore re-formulate well control optimisation as a decision support problem. The aim of this paper is to develop accurate emulators for the expected NPV objective function by exploiting known simulator behaviour in order to efficiently perform decision support.

2.2. OLYMPUS Exploratory Analysis. We first construct a maximin Latin hypercube design and run a wave 0 of $n = 20$ exploratory simulations using all $N = 50$ OLYMPUS models. An important feature is the adherence of OLYMPUS simulations to target production and injection rate decision parameters. For one vector of decision parameter settings Figure 1 compares the input target control rates (black dashed lines) with the corresponding outputted achieved rates over the 50 ensemble members (coloured traces). It is immediately evident in all plots that the input targets are not strictly adhered to for the full duration of the control intervals; a consequence of the underlying physics programmed into the OLYMPUS model, including constraints on BHP, resulting in such deviations between the actual and targeted control values. It is this behaviour which motivates our structured emulation approach described and demonstrated in section 7. Another interesting facet with potential ramifications for emulation and decision support is the vastly different relative absolute contributions of oil and water to the NPV objective function. An assessment is shown in Figure 12, along with further discussion in Appendix A.1.

3. Bayesian Emulation.

3.1. Methodology. An emulator is a stochastic belief specification for a deterministic or stochastic function that provides a fast and efficient statistical approximation, yielding predictions for as yet unevaluated parameter settings, along with a corresponding statement of the associated uncertainty [9, 57, 11]. They are frequently employed for computationally expensive simulators across a range of scientific and industrial applications to perform tasks



(a) WOPR.PROD.2.

(b) WOPR.PROD.10.

(c) WWIR.INJ.2.

(d) WWIR.INJ.3.

Figure 1: Comparison of the target and achieved control rate decision parameters for an OLYMPUS wave 0 exploratory simulation over the full multi-model ensemble. Plots show the OLYMPUS output actual control rates time series as coloured lines for each ensemble member for the Well Oil Production Rates (WOPR, top) or Well Water Injection Rates (WWIR, bottom) for the four wells within the CWG. The black dashed lines show the corresponding decision parameters of target production or injection rates respectively. The differences between the output traces and the inputs highlights the deviations between the target control and what is achieved due to physical constraints applied in the model.

including: calibration [33, 28]; history matching [10, 11, 57, 14]; uncertainty quantifications [38]; sensitivity analyses [36, 32]; and decision support [35, 64].

For a computer model $f(\mathbf{d})$ the i^{th} univariate output is denoted by the function $f_i(\mathbf{d})$,

where $\mathbf{d} \in \Omega \subset \mathcal{R}^D$ is a vector of (decision) parameters in space Ω. We employ Bayesian emulators of the general form in (3.1) [11, 57, 59]:

$$\begin{aligned} f_i(\mathbf{d}) &= \mathbf{g}_i(\mathbf{d}_{A_i})^{\mathrm{T}}\boldsymbol{\beta}_i + u_i(\mathbf{d}_{A_i}) + w_i(\mathbf{d}) \\ &= \sum_{j=1}^{p} \beta_{ij} g_{ij}(\mathbf{d}_{A_i}) + u_i(\mathbf{d}_{A_i}) + w_i(\mathbf{d}) \end{aligned} \tag{3.1}$$

The subscript A_i denotes a subset of active inputs which are the parameters deemed to be most influential for $f_i(\mathbf{d})$, where $|A_i| = D' \leq D$. Within the emulator the first term models the global function behaviour of $f_i(\mathbf{d})$ where the $g_{ij}(\cdot)$ are deterministic functions of the active inputs with unknown scalar regression coefficients, β_{ij} for $j = 1, \ldots, p$, where $p \in \mathbb{N}$. Collectively, these are denoted by the vector function $\mathbf{g}_i^{\mathrm{T}}(\cdot) = \begin{pmatrix} g_{i1}(\cdot) & \cdots & g_{ip}(\cdot) \end{pmatrix}$, and the vector $\boldsymbol{\beta}_i^{\mathrm{T}} = \begin{pmatrix} \beta_{i1} & \cdots & \beta_{ip} \end{pmatrix} \in \mathbb{R}^p$ respectively. The second term, $u_i(\cdot)$, models the local behaviour of $f_i(\mathbf{d})$ and is a weakly stationary stochastic process with zero mean and a pre-specified covariance structure. A common choice is the squared exponential covariance function in (3.2) [57, 47], where $\sigma_{u_i}^2$ is a variance hyperparameter, and $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iD'})$ is a D'-vector of (distinct) correlation lengths.

$$\mathrm{Cov}[u_i(\mathbf{d}_{A_i}), u_i(\mathbf{d}'_{A_i})] = \sigma_{u_i}^2 \exp\left\{ -\sum_{k=1}^{D'} \left(\frac{d_{A_i,k} - d'_{A_i,k}}{\theta_{ik}} \right)^2 \right\} \tag{3.2}$$

The third term in (3.1), $w_i(\mathbf{x})$, is an uncorrelated, zero-mean nugget term with covariance:

$$\mathrm{Cov}[w_i(\mathbf{d}), w_i(\mathbf{d}')] = \sigma_{w_i}^2 \mathbb{1}_{\{\mathbf{d}=\mathbf{d}'\}} \tag{3.3}$$

This is a white noise process which is included to account for the inactive variables [13, 57] and ensure numerical stability [33]. Further arguments for the inclusion of a nugget term are presented in [2, 24].

We follow a Bayes linear paradigm following the foundations of De Finetti [16, 17] using expectation as a primitive within a second-order belief specification. Moreover, we subscribe to subjective Bayesianism to provide a coherent framework to structure and combine expert prior beliefs with observed data to achieve posterior inferences [19]. Bayes linear methods have numerous advantages including: quick and simple elicitation of subjective prior beliefs; computational tractability; and robust inferences by removing the specification of full prior probability distributions, with Bayes linear emulators having been successfully implemented across numerous applications [11, 57, 59, 21, 60]. An in-depth discussion of Bayes linear statistics can be found in [22], with shorter summaries presented in [18, 20]. Given a design $\mathcal{D} = \{\mathbf{d}^{(1)}, \ldots, \mathbf{d}^{(n)}\}$ and computer model evaluations for output i, $\mathbf{F}_i = \{f_i(\mathbf{d}^{(1)}), \ldots, f_i(\mathbf{d}^{(n)})\}$, the Bayes linear adjusted expectation, variance, and covariance are:

$$\mathbb{E}_{\mathbf{F}_i}[f_i(\mathbf{d})] = \mathbb{E}[f_i(\mathbf{d})] + \mathrm{Cov}[f_i(\mathbf{d}), \mathbf{F}_i]\,\mathrm{Var}[\mathbf{F}_i]^{-1}(\mathbf{F}_i - \mathbb{E}[\mathbf{F}_i]) \tag{3.4}$$

$$\mathrm{Var}_{\mathbf{F}_i}[f_i(\mathbf{d})] = \mathrm{Var}[f_i(\mathbf{d})] - \mathrm{Cov}[f_i(\mathbf{d}), \mathbf{F}_i]\,\mathrm{Var}[\mathbf{F}_i]^{-1}\,\mathrm{Cov}[\mathbf{F}_i, f_i(\mathbf{d})] \tag{3.5}$$

$$\mathrm{Cov}_{\mathbf{F}_i}[f_i(\mathbf{d}), f_i(\mathbf{d}')] = \mathrm{Cov}[f_i(\mathbf{d}), f_i(\mathbf{d}')] - \mathrm{Cov}[f_i(\mathbf{d}), \mathbf{F}_i]\,\mathrm{Var}[\mathbf{F}_i]^{-1}\,\mathrm{Cov}[\mathbf{F}_i, f_i(\mathbf{d}')] \tag{3.6}$$

Full derivation of the Bayes linear emulator adjustment formulae is presented in [40, Sec. 2.4.5]. An alternative full Bayesian approach is Gaussian Process (GP) emulators, as discussed in [33]. Under the Bayes linear formulation we opt for a hyperparameter plug-in approach where they are specified a priori, utilising expert elicitation, before validating using emulator diagnostic techniques [7], as performed in [10, 57, 59].

3.2. Bayes Linear Emulation of the Expected NPV. Bayes linear emulation is directly applied to the expected NPV to explore the 32-dimensional wave 1 decision parameter space utilising the above design and linear model predictions for the ensemble mean NPV for training and validation. This serves as a comparison with our proposed hierarchical emulation approach in subsection 9.3. Following the methodology summarised in section 3, an emulator with a nugget term (see (3.1)) is employed with $f(\mathbf{d}) = U(\mathbf{d}) = \mathbb{E}[\text{NPV}](\mathbf{d})$. Investigations using linear modelling, stepwise selection with the AIC criterion, and with all parameters transformed onto $[-1, 1]$, as in [57], yields a subset of 12 active decision parameters, $\mathbf{d}_{A_{\text{BL}}}$ (this includes all 8 target production rates for producer well 2, and target production rates for producer well 10 for the control intervals starting 1$^{\text{st}}$ January 2016, 2018, 2020, and 2026), and a suggested second-order polynomial mean function form:

$$\mathbb{E}[\text{NPV}](\mathbf{d}_{A_{\text{BL}}}) = \beta_0 + \sum_{d_i \in A_{\text{BL}}} \{\beta_{i,1} d_i + \beta_{i,2} d_i^2\} + \varepsilon \tag{3.7}$$

The residual uncertainty is captured through ε which possesses an estimate residual standard error σ_{lm}. The unknown regression coefficients are assumed to have prior expectation $\boldsymbol{\mu}_\beta = \mathbf{0}$ and an infinite prior uncertainty, with emulator updates exploiting limiting results as $\text{Var}[\boldsymbol{\beta}] \to \infty$ for which formulae are presented in [42, Sec. 2.4.5].

For the residual process it is assumed that $\mathbb{E}[u(\mathbf{d}_{A_{\text{BL}}})] = 0$ and $\mathbb{E}[w(\mathbf{d})] = 0$ with a squared exponential covariance structure ((3.2)) using a single common correlation length hyperparameter. Following the substitution approach for the hyperparameters: $\sigma_u^2 = (1-\rho)\sigma_{lm}^2$ and $\sigma_w^2 = \rho\sigma_{lm}^2$ where $\rho = 0.05$; whilst the correlation length parameter is set to half of the parameter range, hence $\theta = 1$. These choices are validated via emulator diagnostics discussed below. Bayes linear emulator adjustment is performed using (3.4)–(3.6).

Leave-one-out diagnostics suggest that the emulator fits well across the decision space, as shown in Figure 2 of the adjusted expectation with an approximate 95% credible interval of width 3 adjusted standard deviations (following Pukelsheim's 3-sigma rule [44]) versus the expected NPV. 691 of the 702 (98.4%) credible intervals contain the simulated expected NPV, as highlighted by the red dashed line representing equality. Moreover, if we instead employed a Gaussian process emulator, then the 95% credible intervals contain 679 of the simulated expected NPVs; a 96.7% coverage. It is noted that the few cases where these diagnostics are not satisfied tend to yield over-prediction. With a view to decision support this is less of a concern as these regions will not be incorrectly ruled out due to low expected NPVs, while iterative refinement enables more accurate emulation at later waves.

4. Subsampling from Multi-Model Ensembles. Multi-model ensembles are frequently employed to characterise various forms of uncertainty. For example, multi-model ensembles are particularly prevalent in climate science such as in [53, 34, 26] where they are used to represent uncertainty pertaining to differing choices of aspects such as model structure, encoding of

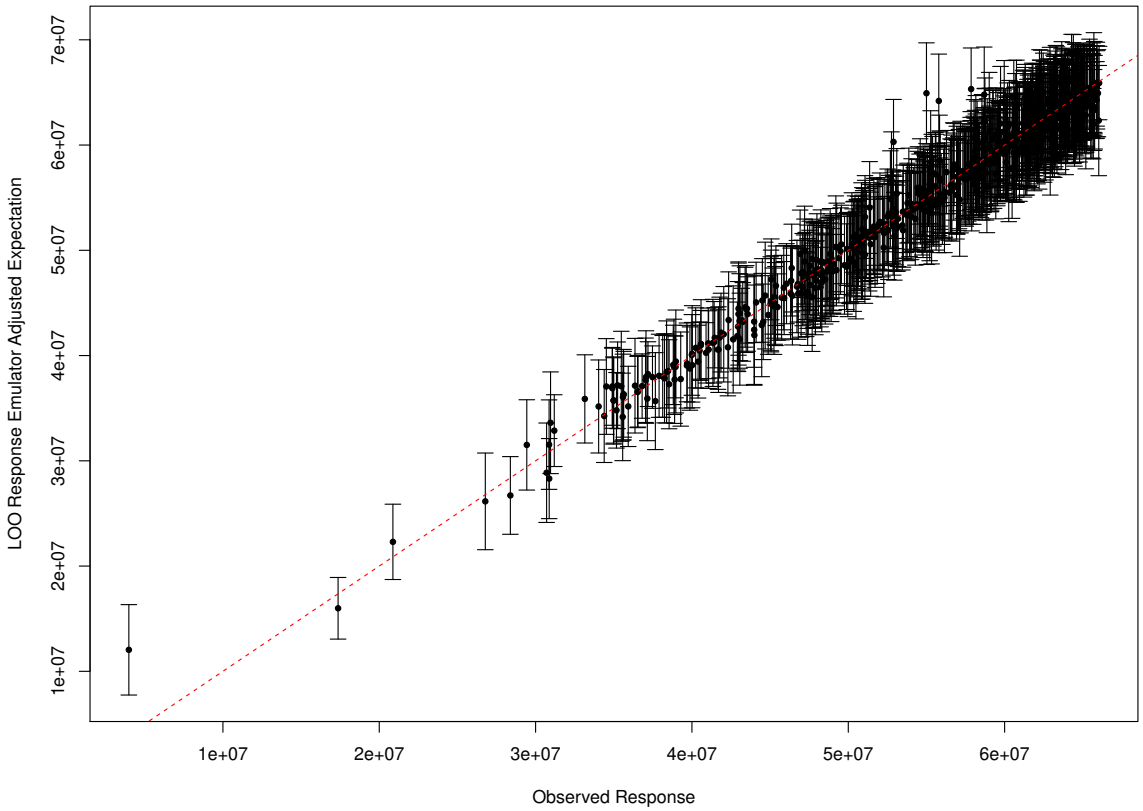


Figure 2: Bayes linear emulator for the expected NPV leave-one-out diagnostics plot showing the emulator adjusted expectations with 95% 3 adjusted standard deviation credible interval error bars versus the simulated expected NPV. The red dashed line denotes equality of the emulator prediction and the simulator output.

physical laws, discretisation scheme, and numerical solvers, in simulating parts of the earth system. In the petroleum industry these are prevalent for representing uncertainty induced by the unknown underlying field geology, reflecting the geologist's beliefs. In scenarios where these are sampled stochastically from an underlying geology model it may be appropriate to employ emulation methods for stochastic computer models such as stochastic Kriging [5], quantile Kriging [46, 43], and heteroscedastic Gaussian Process (hetGP) emulation [8]. For a comprehensive review of stochastic model emulation methodology see [6] and the references therein. In the TNO OLYMPUS Challenge the multi-model ensemble consists of 50 versions of the OLYMPUS model realised from an underlying stochastic geology model. However, this model was not released and so no further realisations were possible, hence the setup of our motivating application is unsuited to stochastic emulation approaches.

Running all 50 models to obtain a representation of geological uncertainty requires a greater number of simulations placing a higher strain on computational resources. In many analyses the multi-model outputs are amalgamated such as through averaging, termed the ensemble mean. This is the case in the TNO OLYMPUS Challenge where the ensemble mean NPV is the focus. Whilst such quantities are easier to analyse and use, the averaging process reduces the benefits of starting with an ensemble by collapsing the uncertainty onto a single value. It is therefore desirable to establish a small collection of models to use as a surrogate, whilst acknowledging any reduction in information gained from the simulations through an appropriate quantification of the uncertainty, and thus develop a multi-model ensemble subsampling technique. Note that a superior choice encompassing uncertainties in the ensemble construction process would be to use an expected utility function over all possible ensembles. However in the TNO OLYMPUS Challenge it is stipulated that the expected NPV objective function is approximated by the ensemble mean NPV, hence for the purpose of this analysis we adhere to this choice, noting the critique of this choice in [40, sec. 3.1] and [42].

4.1. Methodology. The process of identifying a representative subset requires a small exploratory design using all models in the ensemble; a wave 0 design, for example, constructed using a maximin Latin hypercube design [50, 51], in order to assess how well a given subset of models represents the ensemble mean for certain key outputs. First we propose an initial graphical investigation using plots of ensemble mean outputs versus that obtained using individual models. This provides insight into patterns where a strong linear correlation indicates that an individual model may be a good representative for the ensemble mean. Note that such plots are unable to capture the interaction between multiple models' outputs, thus missing where two or more models are jointly able to characterise the ensemble mean, often to a better extent than any one individual model. For large ensembles, this can be a useful screening process to identify a preliminary ensemble subset for further analysis.

Linear models provide a fast and effective tool for predicting the ensemble mean output, $\bar{f}(\cdot)$, for example, $\bar{f}(\mathbf{d}) = \overline{\text{NPV}}(\mathbf{d})$, from individual model outputs, $f^{(i_k)}(\cdot)$, $i_k \in \{1, \ldots, N\}$ distinct, as well as for quantifying the induced uncertainty. For an ensemble of size N the aim is to select the best subset of $\tilde{N} < N$ (to be determined) models. Since the ensemble mean is a linear combination of the individual models' output, an affine linear transformation of a subset of models is expected to yield good approximations. We propose the linear model in (4.1) where α_{ES} and $\beta_{k,\text{ES}}$ are unknown regression coefficients to be estimated, and $\varepsilon_{\text{ES}}(\mathbf{d})$ is an uncorrelated error term. Here ES refers to "Ensemble Subsampling".

$$\bar{f}(\mathbf{d}) = \alpha_{\text{ES}} + \sum_{k=1}^{\tilde{N}} \beta_{k,\text{ES}} f^{(i_k)}(\mathbf{d}) + \varepsilon_{\text{ES}}(\mathbf{d}) \tag{4.1}$$

Depending on the size of N, either an exhaustive or stepwise model selection may be performed to identify the most suitable choice of $\tilde{N}$ and model subset, for example, using the Akaike Information Criterion (AIC) or Bayesian information Criterion (BIC). Note that at later stages within an analysis it is always possible to modify this choice if increased accuracy is required; a scenario that naturally occurs within iterative procedures such as history matching and decision support. In the context of petroleum reservoir field development optimisation we

refer to this as Efficient Geological Ensemble Subsampling (EGES) [42]. Using a subset of models does result in some information loss compared with simulating from the full ensemble. However, the described statistical approach provides a quantification of the additional induced uncertainty as a consequence of using fewer models. In setups where the primary focus is the ensemble mean output the treatment of ensemble variability has been collapsed to a single number and thus neglected. Moreover, running fewer models enables greater exploration of the parameter space and hence presents opportunities to address other sources of uncertainty such as parametric uncertainty. This technique is related to second-order multi-model ensemble exchangeability in [48] where coexchangeability is used to establish a link between: the output of individual models; a common "representative simulator", which we interpret as the ensemble mean simulator; the output for the real world system as the true expected NPV with respect to all possible geological configurations; as well as any system observations.

4.2. Subsampling from a Geological Multi-Model Ensemble Results. At the EAGE/ TNO Workshop on OLYMPUS Field Development Optimization [15] a number of participant teams employed ensemble sub-setting techniques on the 50 Olympus models to aid computational tractability of their chosen optimisation procedure for obtaining a well control strategy. This included: selecting a single representative model in [39], or based on geological modelling insight, specifically using the net hydrocarbon thickness map for upper and lower layers of the reservoir [25]; a risk averse approach to optimisation using the 4 worst performing ensemble members according to the Conditional Value at Risk (CVaR) criterion evaluated for the TNO defined base strategy [52]; and a stochastic rank-based realisation selection process used within a evolutionary strategy optimisation algorithm to produce a different subset at each step of the algorithm [1]. The described Efficient Geological Ensemble Subsampling technique provides a principled approach to selecting ensemble members which also makes efficient use of available computational resources, captures some of the ensemble variability, and not limited to potentially non-robust choices made using the base strategy.

A preliminary graphical investigation is performed using plots of the ensemble mean versus individual models for a range of outputs of interest for the $n = 20$ wave 0 exploratory simulations described in subsection 2.2. Examples are provided in Figure 13 (see Appendix A.2). It is unnecessary to sub-select models exhibiting close individual model output relations with the ensemble mean. Instead we screen for cases where the relationship is easy to model, for example, with a preference for linear associations with small output variation, identifying an initial set of 9 OLYMPUS models for further exploration via linear modelling. Further discussion is found in Appendix A.2, and in [40, Sec. 4.2].

In order to capture the interacting effects of the different OLYMPUS models, the subsampling technique utilising the linear model in (4.1) is implemented. First it is applied to the proposed OLYMPUS subset before extending to all models via both directions stepwise selection starting from the full model and using AIC as the model selection criterion. It is established that a subset of only $\tilde{N} = 3$ models is sufficient for a large number of the investigated outputs, yielding high Adjusted R^2 values shown in Figure 14 in Appendix A.2. The optimal collection for the ensemble mean NPV is OLYMPUS 25, 33, & 45. The fitted linear model provides an efficient means of prediction and uncertainty quantification by using only 3 ensemble members yielding substantial computational savings. This is a novel application

of such multi-model ensemble subsampling techniques in petroleum reservoir engineering.

For computational reasons only 20 runs were available for this pilot study. Discussions with petroleum reservoir engineers suggested that these 20 runs would be sufficient for the purpose of identifying a representative subset since the complexity of this part of the model was not expected to be too high. It is therefore possible to reliably select 3 of the 50 OLYMPUS models based off this collection of simulations, as well as fit linear models to ensemble mean quantities of interest and use these over unexplored regions of the parameter space. If more simulations over the full ensemble were available then it is possible to use diagnostics to verify the robustness of this subset choice. As highlighted above, other participants in the TNO OLYMPUS Well Control Optimisation Challenge employed model sub-selection, but via less statistically principled approaches.

5. Targeted Bayesian Design of Simulations.

5.1. Methodology. We propose a targeted Bayesian design algorithm to efficiently sample from the decision parameter space by incorporating prior knowledge of both the parameter range and their time ordered consecutive differences. This is a generalisation of the design approach presented in [42] and [40, Sec. 4.3] where it is tailored towards the well control optimisation problem. Without loss of generality let $\mathbf{d} \in \mathbb{R}^D$ be a time ordered vector of decision parameters, thus they are not mutually independent. A difference constraint stipulates that $|d_i - d_{i-1}| \leq \Delta_i,\ i = 2, \ldots, D$. The decision parameter space is no longer a hypercube, thus motivating the need for a Bayesian design informed by this prior information.

Our targeted Bayesian design algorithm involves a re-parameterisation and sampling the sum of the parameters and their time consecutive differences. First assume each $d_i \in [0, 1]$. Define $t = \sum_{i=1}^{D} d_i \in [0, D]$ to be the sum of the parameters and $\delta_i = \frac{d_i - d_{i-1}}{\sqrt{2}}$ for $i = 2, \ldots, D$, to be the scaled differences where the scaling by $\frac{1}{\sqrt{2}}$ is required due to the rotation of the parameter space in this alternative parametrisation. The new parameters are mutually orthogonal with $t \in [0, D]$ and $|\delta_i| \leq \Delta_i' = \frac{\Delta_i}{\sqrt{2}}$. This re-parametrisation is represented by the linear transformation $(t, \boldsymbol{\delta})^{\mathrm{T}} = L\mathbf{d}$ in (5.1) where L is the transformation matrix. Sampling in the re-parametrised space automatically satisfies the difference constraints with a sample for $\mathbf{d}$ obtained via $\mathbf{d} = L^{-1}(t, \boldsymbol{\delta})^{\mathrm{T}}$. The range constraints, $d_i \in [0, 1]$, must also then be verified.

$$\begin{pmatrix} t \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \vdots \\ \delta_D \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & \cdots & 1 & 1 \\ -1/\sqrt{2} & 1/\sqrt{2} & 0 & 0 & \cdots & 0 & 0 \\ 0 & -1/\sqrt{2} & 1/\sqrt{2} & 0 & \cdots & 0 & 0 \\ 0 & 0 & -1/\sqrt{2} & 1/\sqrt{2} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \\ \vdots \\ d_D \end{pmatrix} \tag{5.1}$$

In order to ensure good exploration along the t-direction, which is thought to be important by petroleum reservoir engineers for the application, we propose preserving an initial sample of the parameter sums, and then uniformly resample $\boldsymbol{\delta}$ until both constraint types are satisfied. There is freedom to choose the sampling distribution of t with probability density function, $f_T(t)$, dependent on the analysis. We use a truncated normal distribution with other examples being: uniform; mixture of uniforms; and transformed beta distributions. Orthogonal

projection of the samples onto the line $d_1 = d_2 = \cdots = d_D$ will approximately follow the specified distribution.

The process of generating a sample of size n for an independent subgroup of parameters is described in the rejection style Algorithm 5.1 yielding matrix B in which each column is a sampled vector of decision parameters. If t is close to 0 or D the rejection step can be computationally time consuming with the efficiency improved by sampling the differences conditional on t. This algorithm may be applied separately to each independent subgroup of decision parameters for improved efficiency before combing to obtain a design over the full decision parameter space. Design optimisation may be performed using standard design selection criteria, for example, minimax or maximin design [51], both for the designs by subgroup and the full design by combining using random permutations. Further discussion of the targeted Bayesian design method is presented in [40, Sec. 4.3].

Algorithm 5.1 Sampling of parameter sums and differences preserving the initial sum of parameters sample.

Let $\mathbf{t}$ be a vector of n samples of $t \sim f_T(t)$
Let B be an empty matrix of D rows
Define $dimension(\cdot)$ to be a function to obtain the length of a vector
while $dimension(\mathbf{t}) > 0$ **do**
  for all t in **t do**
    Set $\epsilon_i = \min\{t, D - t, \Delta_i'\}$
    Generate $\delta_i \mid t \sim \mathcal{U}[-\epsilon_i, \epsilon_i]$, for $i = 2, \ldots, D$
  end for
  Row bind $\mathbf{t}, \boldsymbol{\delta}_2, \ldots, \boldsymbol{\delta}_D$ to form matrix $B_{r,\text{prop}}$
  Compute $B_{\text{prop}} = L^{-1} B_{r,\text{prop}}$
  for all Column in B_{prop} **do**
    if Range *conditions* of parameter vector are satisfied **then**
      Join *Column* to B
      Remove corresponding t from $\mathbf{t}$
    else
      Discard *Column*
    end if
  end for
end while
return Matrix B of columns of sampled parameter vectors

5.2. Targeted Bayesian Design of Simulations Results. We employ the targeted Bayesian design methodology from subsection 5.1 to perform decision support through targeted sampling based on prior beliefs of experienced petroleum reservoir engineers regarding the location of optimal decision parameter settings, as well as imposing practical and physical constraints. Firstly, TNO stipulate operational range constraints on the control parameters to be $[0, 900]\text{m}^3/\text{day}$ and $[0, 1600]\text{m}^3/\text{day}$ for production and injection rates respectively, leading to a 32-dimensional hypercube parameter space. Oil reservoir engineers deem large temporal

variation in controls to be unphysical and poor practice, thus suggesting a difference constraint between time consecutive controls. Here we use the notation d_{jk,t_i} for the decision parameters where the index i is replaced by the indices tuple (jk, t_i), where $j \in \{P, I\}$ refers to the well type (P producer, I injector), k is the well number, and t_i is the control interval start date. The 32 decision inputs naturally split into four independent subgroups by well with difference constraints $|d_{jk,t_i} - d_{jk,t_i-1}| \leq \Delta_i,\ i = 2, \ldots, D^{(jk)}$, where $D^{(jk)} = 8$ is the number of control intervals for well jk. A conservative choice is that the maximum permitted change over a two year time interval is $\Delta_i = \frac{1}{3}$ of the operational range for the well type. Consequently the decision parameter space is no longer a hypercube with a volume of 3.45% of the initial hypercube due to the range constraints only.

The targeted Bayesian design algorithm is implemented to generate a $n = 700$ point design. First, for each of the four subgroups of eight decision parameters the normalised parameter sums are sampled from a truncated normal distribution in order to facilitate the exploration of more extreme values of the total sums of the eight normalised parameters than would be the case using a standard uniform or Latin hypercube design. This is perceived to be important based on reservoir engineering insight. Next, the differences are sampled according to the specified value of Δ_i before imposing the operational range constraints (after transforming the normalised parameters to their physical values). Each parameter subgroup and the overall design are approximately optimised with respect to the minimax design selection criterion by comparing candidate designs to a large $20,000$ point uniform random sample (over the constrained parameter space) [51]. Moreover, the optimised design is augmented to include two further decision parameter vectors with all parameters set to either their minimum or maximum values since it is of interest to observe the model behaviour at these extremes.

The final wave 1 design for eight producer well 2 parameters is illustrated in Figure 3. The plots next to the diagonal highlights the difference constraints as points are clustered between two clearly defined diagonal parallel bounds. Since the final two control intervals are of length 4 years a greater change of up to $\frac{2}{3}$ of the parameter operational range is permitted, hence the wider bands. In addition, the difference constraints affect decision parameters at larger time separations where there are fewer points away from the diagonal, although is less pronounced for greater time gaps. This design is evaluated for the identified subset of 3 OLYMPUS models with the linear model used to predict the ensemble mean NPV for which points are coloured green, yellow and red for high, moderate and low NPVs respectively in Figure 3. Note that the presented emulation methodology also works with more traditional space filling designs. Employment of a targeted Bayesian design is to enhance the overall decision support aim, and to incorporate expert knowledge regarding the reservoir behaviour and practical decision implementation.

6. Divide-and-Conquer Approach.

6.1. Methodology. Outputs of interest often consist of linear combinations of computer model outputs, as is the case in the TNO OLYMPUS Well Control Optimisation Challenge where the focus is the ensemble mean NPV objective function. One approach is to directly emulate this output using methodology such as that described in subsection 3.1. However, this ignores the potential gains which can be achieved by decomposing these sums into their constituent simulation outputs and instead emulating each of these before recombining. We

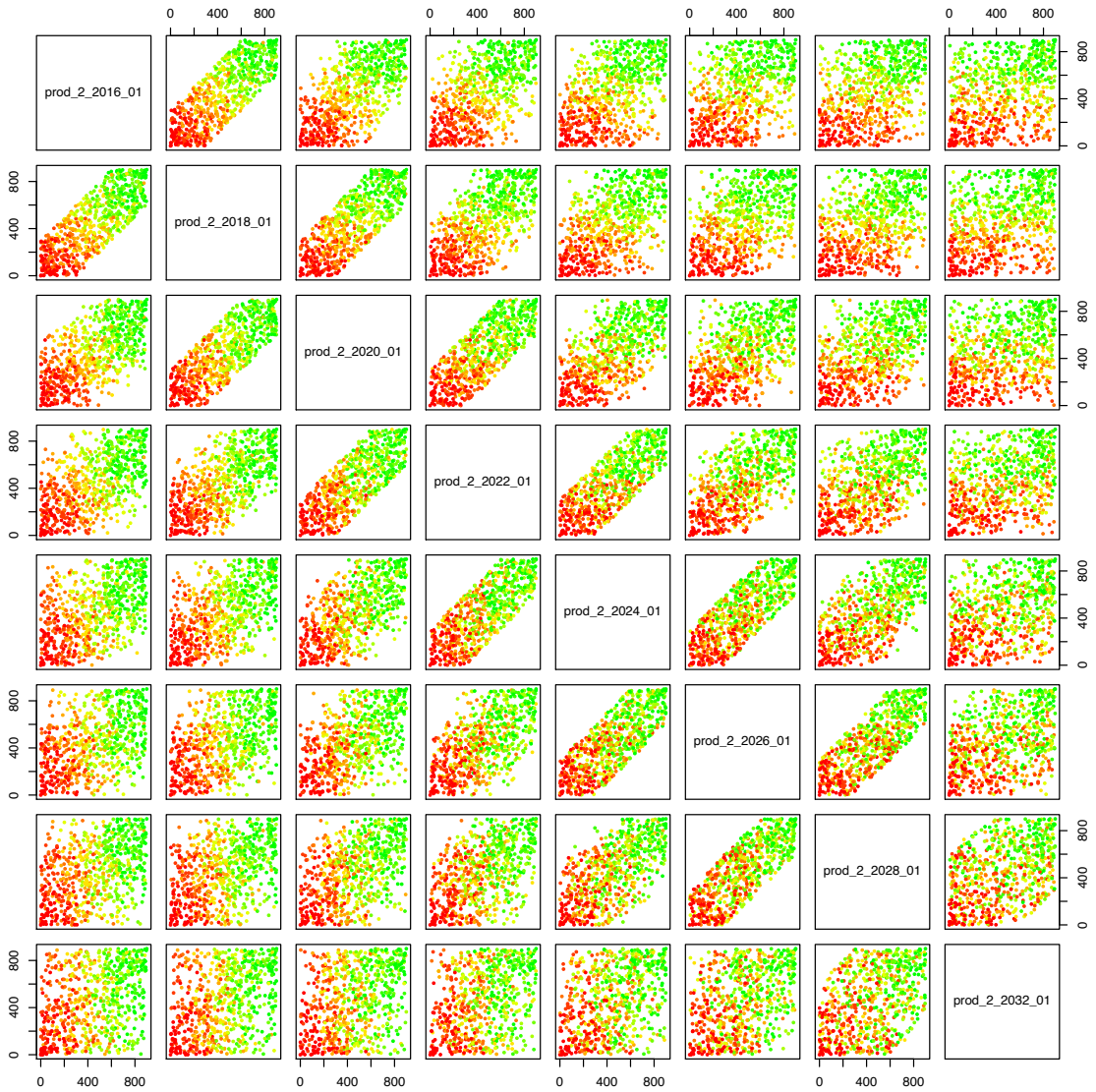


Figure 3: Wave 1 702 point design OLYMPUS producer well 2 decision parameters pairs plots. Points are coloured by the multi-model ensemble subsampling linear model predicted NPV where green, yellow and red correspond to high, moderate and low NPVs respectively.

term this the "divide-and-conquer" approach. In full generality assume an output $f(\mathbf{d})$ can be expressed as:

$$f(\mathbf{d}) = \sum_{i=1}^{q} a_i f_i(\mathbf{d}) \tag{6.1}$$

where each $f_i(\mathbf{d})$, $i = 1, \ldots, q$, is a constituent simulation output. The emulator update equations for the expectation and variance are:

$$\mathbb{E}_{\mathbf{F}}[f(\mathbf{d})] = \sum_{i=1}^{q} a_i \, \mathbb{E}_{\mathbf{F}_i}[f_i(\mathbf{d})] \tag{6.2}$$

$$\mathrm{Var}_{\mathbf{F}}[f(\mathbf{d})] = \sum_{i=1}^{q} a_i^2 \, \mathrm{Var}_{\mathbf{F}_i}[f_i(\mathbf{d})] \tag{6.3}$$

where $\mathbf{F}_i = \{f_i(\mathbf{d}^{(1)}), \ldots, f_i(\mathbf{d}^{(n)})\}$, and $\mathbf{F} = \{\mathbf{F}_i\}_{i=1}^{q}$. Note that (6.3) it is assumed that independent emulators are fitted for each $f_i(\mathbf{d})$, although this naturally extends to where multivariate emulators are employed by introducing the relevant covariance terms.

6.2. Results. The divide-and-conquer approach is applied in our analysis of the TNO OLYMPUS Well Control Optimisation Challenge. First we emulate the NPV for an individual OLYMPUS model, denoted here by $f(\mathbf{d})$, and defined in (2.1) and (2.3), and omitting the OLYMPUS model index j for clarity of notation. This is a linear combination of the well contributions with weights determined by the associated cost parameters and the discount factor. A natural decomposition is thus:

$$f(\mathbf{d}) = \sum_{i=1}^{8} a_i \left\{ r_{op} \left(\sum_{k \in \{2,10\}} f^{op}_{Pk,t_i}(\mathbf{d}) \right) - r_{wp} \left(\sum_{k \in \{2,3\}} f^{wp}_{Ik,t_i}(\mathbf{d}) \right) - r_{wi} \left(\sum_{k \in \{2,3\}} f^{wi}_{Ik,t_i}(\mathbf{d}) \right) \right\} \tag{6.4}$$

where $f^{op}_{Pk,t_i}(\mathbf{d})$, $f^{wp}_{Ik,t_i}(\mathbf{d})$, and $f^{wi}_{Ik,t_i}(\mathbf{d})$ are the Well Oil Production Total (WOPT), Well Water Production Total (WWPT), and Well Water Injection Total (WWIT) within the 8 control intervals ending at time t_i respectively. The index P and I refer to producer and injector wells respectively, and k is the well number over the set of wells used in this analysis. The coefficients a_i are average discounting factors computed as described in subsection 8.2. It is these 48 constituents which are to be emulated.

In principle we could employ this approach for all ensemble members which comes at the cost of requiring simulations from all models. Within this application it is unnecessary due to the geological multi-model ensemble subsampling performed in subsection 4.2, hence this process is only performed for the NPV for OLYMPUS 25, 33, & 48. The importance of the divide-and-conquer approach will become evident in 7.2 where we exploit known behavioural structure in the constituent outputs to obtain more accurate emulators for the WOPT and WWIT outputs compared with a "black-box" approach.

7. Structured Emulators Exploiting Known Simulator Behaviour.

7.1. Methodology. The emulation methodological development presented here is motivated by the partially known behavioural form of the WOPT and WWIT outputs within control intervals with respect to their corresponding decision input parameters, the target production or injection rate for the control interval respectively, as shown in Figure 4. Within (2.3) for OLYMPUS model j the WOPT and WWIT are the by well constituents of the respective field totals $Q_{j,op}(\mathbf{d}, t_i)$ and $Q_{j,wi}(\mathbf{d}, t_i)$. For small values of the decision parameter

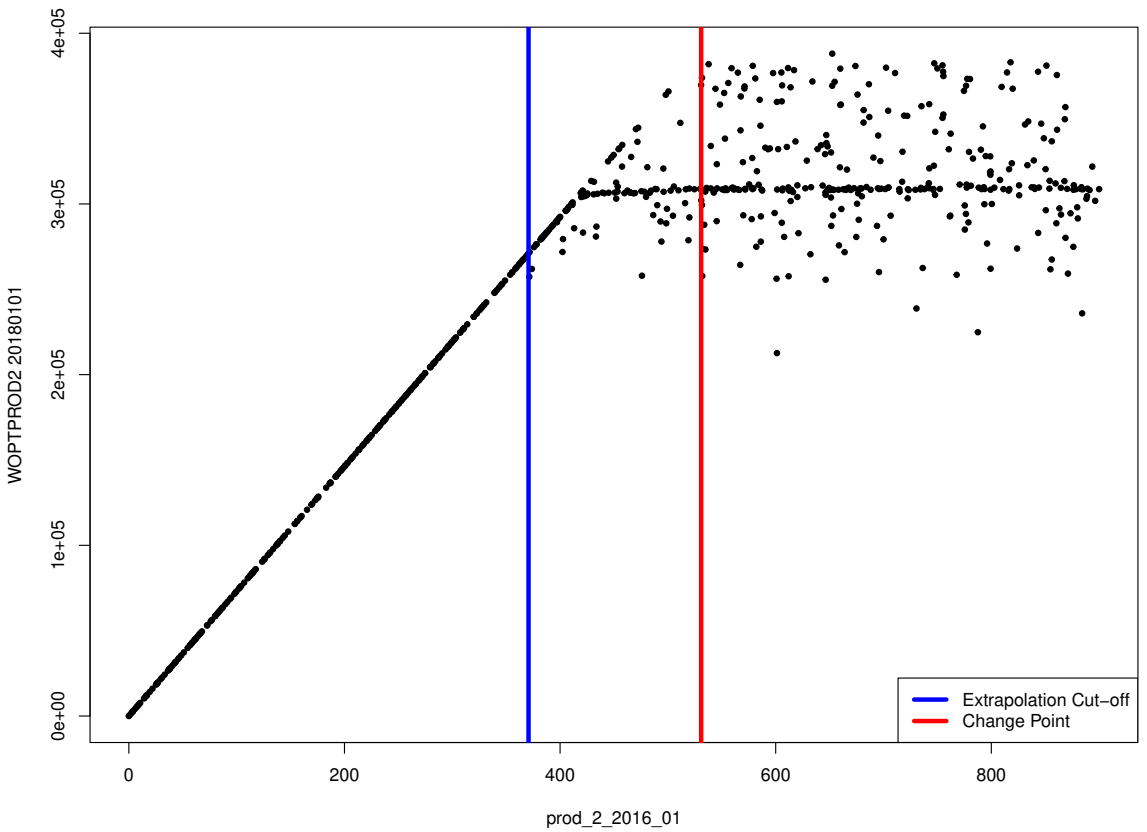


Figure 4: OLYMPUS 25 WOPT output for producer well 2 during the first two years (ending 01/01/2018) versus the corresponding decision input parameter, the target production rate prod_2_2016_01. For small values of prod_2_2016_01 the target is achieved resulting in the perfect linear behaviour up to a change point beyond which the WOPT plateaus as a maximum threshold on the production rate is achieved. The vertical blue and red lines denote the extrapolation cut-off and change point upper bounds respectively.

the output is known as a linear function of this input up to a small tolerance. Beyond a certain value of this parameter there is a departure from this linear behaviour before reaching a plateau with output fluctuations depending on variation in the other parameter values.

These distinct function behavioural regimes within different regions of the parameter space should be exploited to achieve more accurate emulators than applying the general Gaussian Process (GP) or Bayes linear emulators. Another option is to employ partition based emulation approaches such as treed Gaussian processes [23] which split the parameter space parallel and perpendicular to the input axes and fits independent GPs within each region. However whilst treed GPs are a flexible model they do not exploit the physical behaviour

that is known to exist in the model. In addition, a further limitation to their accuracy is the number of design points within each partition region which may constitute a small volume of the overall parameter space, particularly important in the often narrow intermediate region for the developed methodology. These considerations have implications during subsequent (decision) analyses utilising the emulator.

Here we present novel methodology which is able to capture both the observed output structure in Figure 4 and an output upper bound which involves splitting the parameter space and output modes of behaviour into three regions: slop, intermediate, and plateau. In this section we denote the computer model output by $f(\mathbf{d})$, with decision inputs $\mathbf{d}$, and assume without loss of generality that the behaviour manifests with respect to decision parameter d_1. In subsection 7.2 we link this notation to the TNO OLYMPUS Challenge application.

7.1.1. Change Points. In the slope region d_1 determines the output according to a known functional relationship up to a small tolerance $\delta \geq 0$ whilst also imposing an upper bound on $f(\mathbf{d})$. Here we focus on a linear map with respect to d_1 with known intercept, α, and gradient, β, noting that the methodology also extends to other known relationships. Within the plateau region this known behaviour is not the case with uncertainty induced by $\mathbf{d}\backslash d_1$. In principle this leads to these two regions only separated by a change point denoted c. However, the value of d_1 of the transition from slope to plateau is unknown, depends on all other decision parameters, and given only a finite number of simulations is impossible to exactly determine. Consequently, in practical application, there are three distinct regions of behaviour: the slope and plateau separated by an additional uncertain region believed to contain the unknown change point. This is labelled as the intermediate region where there is a mixture of adherence to this known (linear) relationship and model output exhibiting uncertainty around the plateau. Given a design $\mathcal{D}$ and simulator output $\mathbf{F} = \{f(\mathbf{d}^{(1)}), \ldots, f(\mathbf{d}^{(n)})\}$, one option it to estimate the mean change point location.

A conservative change point upper bound estimate, c^u, is defined in (7.1), where $f_{\max} = \max_{\mathbf{d}\in\mathcal{D}} f(\mathbf{d})$, and $\delta_u \geq 0$ is a tolerance included for numerical stability and to ensure that an upper bound is obtained.

$$c^u = \min_{d_1}\{d_1 \mid \alpha + d_1 \cdot \beta \geq f_{\max} + \delta_u\} \tag{7.1}$$

This is the smallest value of d_1 such that if simulator output achieved the upper bound, $\alpha + d_1 \cdot \beta$, then this exceeds the largest simulated value (plus a tolerance) over $\mathcal{D}$.

An estimate for the change point lower bound is defined in (7.2), where $f_{\text{diff}}(\mathbf{d}) = \alpha + d_1 \cdot \beta - f(\mathbf{d})$ is the difference between the theoretical maximum and the simulated output, and $\delta_l \geq 0$ is another numerical stability tolerance.

$$\begin{aligned} c^l = \frac{1}{2}\Bigg(& \arg\min_{d_1|\mathbf{d}\in\mathcal{D}}\{f(\mathbf{d}) < \alpha + d_1 \cdot \beta - \delta_l\} \\ & + \arg\max_{d_1|\mathbf{d}\in\mathcal{D}}\left\{d_1 < \arg\min_{d_1|\mathbf{d}\in\mathcal{D}}\{f(\mathbf{d}) < \alpha + d_1 \cdot \beta - \delta_l\}\right\}\Bigg) \end{aligned} \tag{7.2}$$

Figure 5 provides an illustration of the change point lower bound for the output WOPT-PROD2_20180101 with respect to the target rate prod_2_2016_01.

7.1.2. Extrapolation Cut-Offs. The two distinct modes of behaviour suggests an emulator be fitted piecewise using a combination of the more accurate knowledge in the slope region, and the less well understood behaviour in the plateau region. Noting the change point location uncertainty, for the plateau region we propose fitting an emulator based only on data which is almost certainly on the plateau using the change point upper bound estimate. For $f(\cdot)$ this is design points with $d_1 \geq c^u$. In order to connect the slope and plateau regions we must extrapolate the plateau emulator. It is necessary to introduce an extrapolation cut-off, denoted b, beyond which the emulator should not be extrapolated. This is due to limited plateau training data issues. For simulator output $f(\mathbf{d})$ this is defined with respect to the same decision parameter, d_1. The decision space is thus split into three distinct regions:

1. **Slope Region:** where $d_1 < b$;
2. **Intermediate Region:** where $b \leq d_1 < c^u$, for which there is uncertainty as to whether simulator output falls on the slope or in the plateau;
3. **Plateau Region:** where $d_1 \geq c^u$.

There is an estimation trade-off between overly cautious small values which fails to alleviate the above issue, and large values risking points being wrongly classified as on the slope. A suitable and sufficiently conservative approach is to use the change point lower bound, so $b = c^l$ (see (7.2)).

7.1.3. Structured Emulation with Upper Truncation. Prior information for $f(\mathbf{d})$ stipulates that it cannot exceed a maximum (up to a tolerance) determined by d_1 and the coefficients α & β. This upper bound is $\alpha + d_1 \cdot \beta$. Alongside the above parameter space dichotomy, this constraint is imposed through an upper truncation with structured emulators respecting partially known behaviour of these simulator outputs. First a preliminary Bayes linear emulator for $f(\mathbf{d})$ is fitted using a sub-design, $\mathcal{D}' = \{\mathbf{d} \mid \mathbf{d} \in \mathcal{D},\ d_1 > c^u\}$, with corresponding simulator output $\mathbf{F}' = \{f(\mathbf{d}) \mid \mathbf{d} \in \mathcal{D}'\}$. This represents the behaviour in the plateau region. By construction, all parameter settings in $\mathcal{D}'$ do not adhere to the target rate and hence are in the plateau, thus providing reliable information on which to construct this part of the emulator. This preliminary emulator is only evaluated for $\mathbf{d}$ satisfying $d_1 \geq b$ and is compared with the upper bound in a classification step determining the structured emulator form:

1. **Slope Region:** If $d_1 < b$ **or** for the preliminary emulator $\mathbb{E}_{\mathbf{F}'}[f(\mathbf{d})] - 3\sqrt{\mathrm{Var}_{\mathbf{F}'}[f(\mathbf{d})]} > \alpha + d_1 \cdot \beta$, then collapse the emulator such that for the structured emulator $\mathbb{E}_{\mathbf{F}'}[f(\mathbf{d})] = \alpha + d_1 \cdot \beta$ with fixed maximum absolute errors of size δ.
2. **Intermediate Region:** If the preliminary emulator satisfies $\mathbb{E}_{\mathbf{F}'}[f(\mathbf{d})] - 3\sqrt{\mathrm{Var}_{\mathbf{F}'}[f(\mathbf{d})]} \leq \alpha + d_1 \cdot \beta < \mathbb{E}_{\mathbf{F}'}[f(\mathbf{d})] + 3\sqrt{\mathrm{Var}_{\mathbf{F}'}[f(\mathbf{d})]}$, a truncated Gaussian process (truncated GP) emulator is used with mean and variance determined by (7.3) and (7.4) respectively [30].
3. **Plateau Region:** In all other cases where $\mathbb{E}_{\mathbf{F}'}[f(\mathbf{d})] + 3\sqrt{\mathrm{Var}_{\mathbf{F}'}[f(\mathbf{d})]} \leq \alpha + d_1 \cdot \beta$, the preliminary emulator output is used.

Alternative width credible intervals may be used depending on the level of conservativeness desired within an analysis with justification based on the Vysochanskij-Petunin inequality [62].

7.1.4. Structured Emulation with Two-Sided Truncation. An alternative variant of structured emulation uses a two-sided truncation where a lower truncation is also imposed. This may be either a constant γ, such as to enforce that an output is non-negative via $\gamma = 0$, or a function of the parameters $\gamma(\mathbf{d})$. For clarity of notation we denote this by γ. Both upper and lower constraints are utilised alongside the preliminary Bayes linear emulator in a modified classification step:

1. **Slope Region:** If $d_1 < b$ **or** for the preliminary emulator $\mathbb{E}_{\mathbf{F}'}[f(\mathbf{d})] - 3\sqrt{\mathrm{Var}_{\mathbf{F}'}[f(\mathbf{d})]} > \alpha + d_1 \cdot \beta$, then collapse the emulator such that for the structured emulator $\mathbb{E}_{\mathbf{F}'}[f(\mathbf{d})] = \alpha + d_1 \cdot \beta$ with fixed maximum absolute errors of size δ.
2. **Intermediate Region:** As for the upper truncation version, if the preliminary emulator satisfies $\mathbb{E}_{\mathbf{F}'}[f(\mathbf{d})] - 3\sqrt{\mathrm{Var}_{\mathbf{F}'}[f(\mathbf{d})]} \leq \alpha + d_1 \cdot \beta < \mathbb{E}_{\mathbf{F}'}[f(\mathbf{d})] + 3\sqrt{\mathrm{Var}_{\mathbf{F}'}[f(\mathbf{d})]}$, **or** if the additional criterion of $\mathbb{E}_{\mathbf{F}'}[f(\mathbf{d})] - 3\sqrt{\mathrm{Var}_{\mathbf{F}'}[f(\mathbf{d})]} < \gamma$ **whilst** $\mathbb{E}_{\mathbf{F}'}[f(\mathbf{d})] + 3\sqrt{\mathrm{Var}_{\mathbf{F}'}[f(\mathbf{d})]} \leq \alpha + d_1 \cdot \beta$, a truncated GP emulator is evaluated. The mean and variance are determined by (7.3) and (7.4) respectively.
3. **Plateau Region:** In all other cases where $\mathbb{E}_{\mathbf{F}'}[f(\mathbf{d})] - 3\sqrt{\mathrm{Var}_{\mathbf{F}'}[f(\mathbf{d})]} \geq \gamma$ **and** $\mathbb{E}_{\mathbf{F}'}[f(\mathbf{d})] + 3\sqrt{\mathrm{Var}_{\mathbf{F}'}[f(\mathbf{d})]} \leq \alpha + d_1 \cdot \beta$, use the preliminary emulator.

The structured emulation methodology utilises a truncated Gaussian process (truncated GP) emulator for which the mean and variance are determined by (7.3) and (7.4) respectively [30], where $\phi(\cdot)$ and $\Phi(\cdot)$ represent the probability density and cumulative distribution functions respectively of a standard normal distribution. These are computed assuming a preliminary Gaussian process emulator with posterior mean and variance, abbreviated to μ and σ^2 respectively, equal to the computed adjusted expectation and variance, and truncation bounds $p = \gamma$ and $q = \alpha + d_1 \cdot \beta$, with $\nu = \frac{p-\mu}{\sigma}$, and $\omega = \frac{q-\mu}{\sigma}$. This form of emulation is used in the intermediate uncertain region around the change point's true location.

$$\mathbb{E}_{\mathbf{F}'}[f(\mathbf{d}) \mid p < f(\mathbf{d}) < q] = \mu + \sigma \frac{\phi(\nu) - \phi(\omega)}{\Phi(\omega) - \Phi(\nu)} \tag{7.3}$$

$$\mathrm{Var}_{\mathbf{F}'}[f(\mathbf{d}) \mid p < f(\mathbf{d}) < q] = \sigma^2 \left[1 + \frac{\nu\phi(\nu) - \omega\phi(\omega)}{\Phi(\omega) - \Phi(\nu)} - \left(\frac{\phi(\nu) - \phi(\omega)}{\Phi(\omega) - \Phi(\nu)} \right)^2 \right] \tag{7.4}$$

Compared to using standard GP, Bayes linear, or treed GP emulators, the structured approach yields improved accuracy by encapsulating the known output structure and constraints, along with an increase in speed and efficiency; a consequence of using fewer design points in the fitting. Moreover, both the change point upper bound and extrapolation cut-off estimation processes are computationally very cheap, whilst the use of a truncated GP helps reduce the reliance on accurate estimation of the extrapolation cut-off. Another benefit to this approach is that in principle few points are required within the intermediate region between the slope and plateau since the emulator is only trained on those which exceed the change point upper bound, although they do form a useful emulator diagnostic check. Further commentary on the accuracy and speed of structured versus Bayes linear emulators can be found in the application to NPV constituents exhibiting such constrained behaviour in subsection 7.2, as well as in the comparison of employing Bayes linear emulators versus structured emulators within the combined hierarchical emulation framework for the ensemble mean NPV

in subsection 9.3.

7.2. Results. For each OLYMPUS model the NPV is determined by the oil production, water injection, and water production. Following subsection 6.2 we decompose the NPV calculation into WOPT, WWIT, and WWPT, by both model and control interval, as in (6.4). The WOPT and WWIT over a control interval are observed to follow the structured behaviour where the quantity is equal to the corresponding target rate decision parameter multiplied by the length of the time interval up to an unknown change point beyond which there is a plateau in the behaviour. In addition, the value of this decision parameter also imposes a maximum achievable output over this time interval. This is illustrated for the OLYMPUS 25 WOPT for producer well 2 over the first two year control interval (01/01/2016 to 01/01/2018) in Figure 4 which is used as a running example.

The structured emulation with upper truncation methodology is employed separately for each of the WOPT and WWIT within control interval constituents for wells in the CWG. For outputs $f^{op}_{Pk,t_i}(\mathbf{d})$ and $f^{wi}_{Ik,t_i}(\mathbf{d})$ the corresponding decision parameter is the target production or injection rate for this interval and is denoted d_{jk,t_i}, where $j \in \{P, I\}$, k is the well number, and t_i is the control interval end year, which equates to d_1 in subsection 7.1. These target control rates should be adhered to for the entire duration of the interval, Δt_i. However, this is not always possible due to Bottom Hole Pressure (BHP) constraints which results in a departure from the target and the observed different modes of behaviour across the parameter space. The upper truncation is therefore obtained by specifying $\alpha = 0$ and $\beta = \Delta t_i$ throughout subsection 7.2.

Conservative change points upper bounds, c^u_{jk,t_i}, and extrapolation cut-offs, b_{jk,t_i}, are estimated using (7.1) and (7.2) respectively over the wave 1 simulations, with tolerances $\delta_u = \delta_l = 10$ chosen to ensure numerical stability. These are shown for WOPTPROD2_20180101 versus the target rate prod_2_2016_01 at $b_{P2,2016} = c^l_{P2,2016}$ and $c^u_{P2,2016}$ in Figure 4 by the vertical blue and red lines respectively. In addition, the estimation process of the change point lower bound (or extrapolation cut-off) for this output is depicted in Figure 5 where the red line represents the slope and upper bound if the target is adhered to for the entire control interval. The vertical blue line denotes $c^l_{P2,2016}$ as the midpoint between the first simulation decision parameter setting not on the slope; hence with $f_{\text{diff}}(\mathbf{d}) > \delta_l$ (green point; first term in (7.2)), and the decision parameter setting with the largest value of $d_{P2,201601}$ which is less than this first departure point previously obtained (magenta point; second term in (7.2)). Further results are displayed in Figure 15 (in Appendix A.3) showing the regions in which the "true" change points are believed to be situated for all WOPT and WWIT for each of the three wave 1 OLYMPUS models.

For each NPV constituent the next stage is to fit a preliminary Bayes linear emulator where the deterministic functions of the active decision parameters are of the form:

$$m(\mathbf{d}_{A_{jk,t_i}}) = \mathbf{g}(\mathbf{d}_{A_{jk,t_i}})^{\mathrm{T}}\boldsymbol{\beta} = \beta_0 + \sum_{d_i \in A_{jk,t_i}} \{\beta_{i,1} d_i + \beta_{i,2} d_i^2\} \tag{7.5}$$

It is assumed that the active decision parameters comprise all decisions which take place in the past of the output for reasons of temporal consistency. For our running example this is $A_{jk,t_i} = \{\text{prod\_2\_2016\_01}, \text{prod\_10\_2016\_01}, \text{inj\_2\_2016\_01}, \text{inj\_3\_2016\_01}\}$. This is a logical

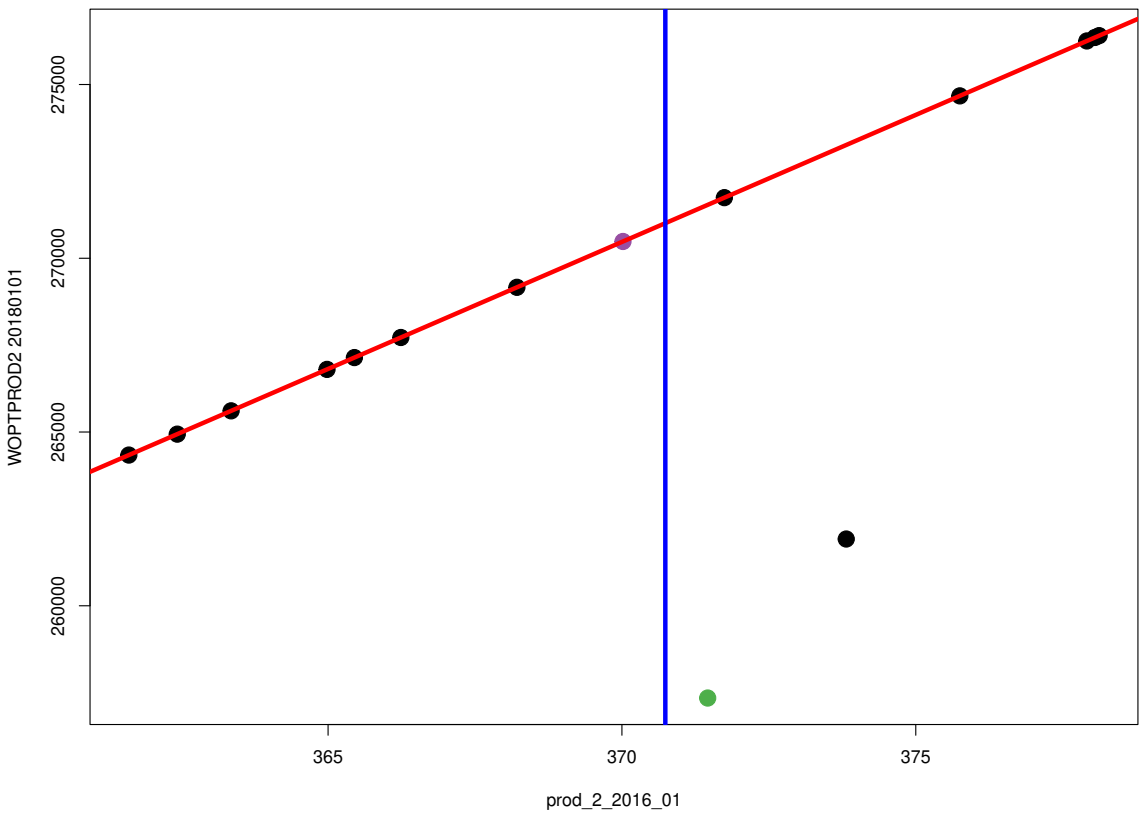


Figure 5: OLYMPUS 25 WOPT for producer well 2 during the first two years (ending 01/01/2018) versus the corresponding target production rate, prod_2_2016_01. The focus is the change point lower bound, $c^l_{P2,2016}$, computed using (7.2), employed as the extrapolation cut-off, and denoted by the vertical blue line. The red line depicts the slope upper bound computed as prod_2_2016_01 $\cdot\, \Delta t_{201801}$ and attained when the target production rate is adhered to for the full control interval. It is shown that $c^l_{P2,2016}$ is the midpoint of the first point not on the slope coloured green, and preceding point to the left of the vertical blue line which is on the slope coloured magenta.

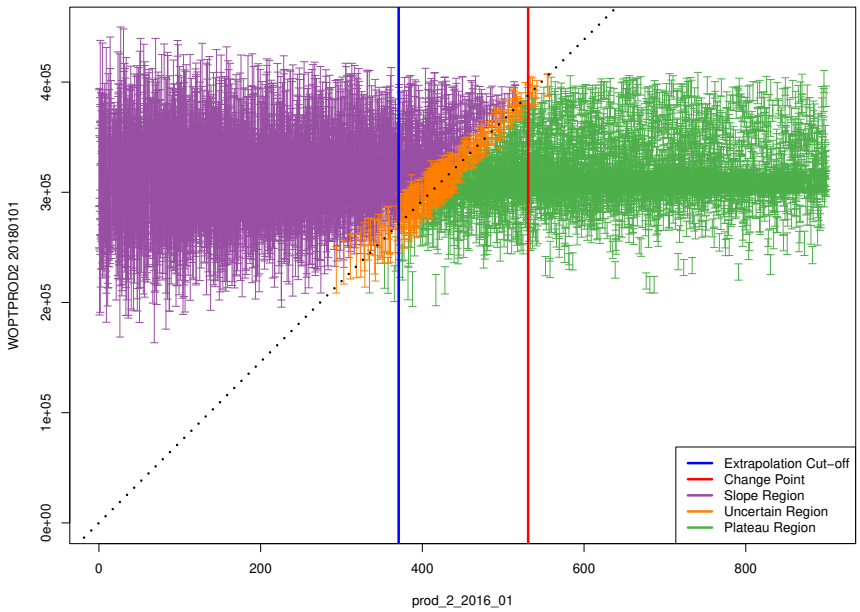
choice since future decisions are physically unable to impact on an output up to the current time, however any past decisions may potentially have an effect. The remainder of each emulator's prior specification is analogous to subsection 3.2, but with the distinction that only those simulation points in $\mathcal{D}' = \{\mathbf{d} \mid \mathbf{d} \in \mathcal{D},\ d_{jk,t_i} > c^u_{jk,t_i}\}$ with output $\mathbf{F}' = \{f(\mathbf{d}) \mid \mathbf{d} \in \mathcal{D}'\}$ are used in the fitting.

For our example of the OLYMPUS 25 WOPT for producer well 2 in the first control interval (ending 01/01/2018) the preliminary emulator predictive 3-sigma credible intervals

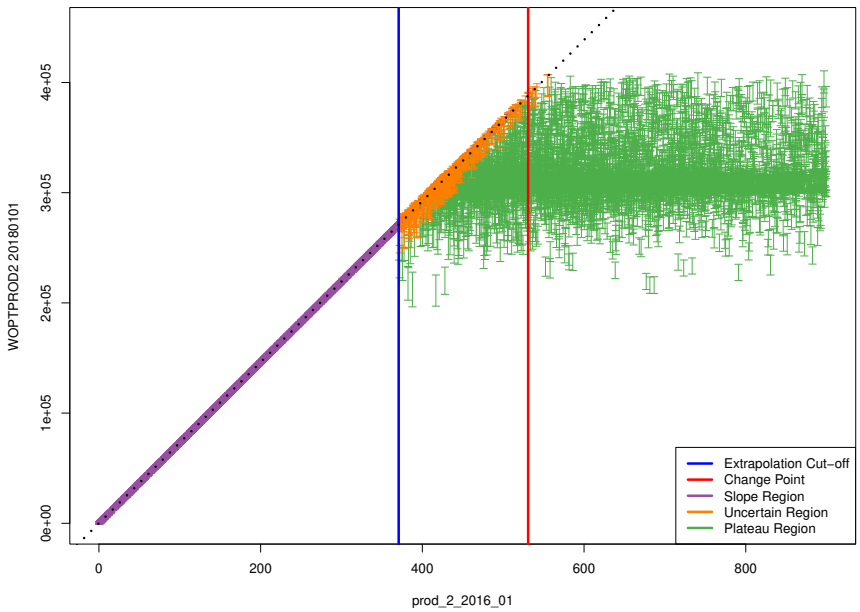
are illustrated in Figure 6a versus the corresponding decision parameter, prod_2_2016_01. This is compared to the theoretical maximum output depicted by the black dotted line determined by the effective target rate for the control interval. The vertical blue and red lines are situated at $b_{P2,2016}$ and $c^u_{P2,2016}$ respectively. The structured emulation methodology using an upper truncation yields the results shown in Figure 6b. Within the slope region, shown in purple, the preliminary emulator credible intervals exceed the constraint and are thus collapsed onto the slope yielding very narrow intervals representing strong beliefs that these decisions will be adhered to for the full control interval. Included are all decision parameter vectors with $d_{P2,2016} < b_{P2,2016}$, as well as cases where $d_{P2,2016} \geq b_{P2,2016}$ in which the preliminary emulator credible interval lower bound exceeds the slope. The uncertain intermediate region, shown in orange, is handled by a truncated GP reflecting the uncertainty in whether the model output is on the slope or relatively close when a target rate is achieved for the majority of a control interval. All of these points lie close to the black dotted slope line with much narrower credible intervals than the preliminary Bayes linear emulator. For the plateau region, shown in green, the preliminary emulator credible interval is well below this slope. It is therefore unnecessary to impose a truncation due to the very small probability that an emulator realisation actually exceeds this physical constraint, hence these intervals are unchanged between the two plots.

Leave-one-out structured emulator diagnostics demonstrate satisfactory results with examples shown in Figure 7 for OLYMPUS 25 NPV constituents WOPT and WWIT in the control intervals ending 01/01/2018 and 01/01/2022 respectively. The first is our running example. Figures 7a and 7c show the emulator adjusted expectation with 95% (3 adjusted standard deviations) credible intervals versus the simulated output. In each case the emulator is exceptionally accurate for smaller NPV constituent values corresponding to where the target rate is adhered to for the entire control interval. For larger simulated outputs believed to be on plateau, the credible interval is wider, whilst the use of a truncated GP emulator for intermediate values demonstrates a reduction in the uncertainty in these locations. The classification step emulator type is best observed in Figures 7b and 7d of the credible intervals versus each NPV constituents' corresponding decision parameter. The structured emulation approach is applied to each of the 3 OLYMPUS models identified in subsection 4.2 for all of the NPV constituents. It is found that the majority of the 95% credible intervals contain the simulated value with the maximum percentage of failures over each output type reported in Table 1. Moreover, no issues are detected in other leave-one-out diagnostic analyses. This demonstrates how incorporating known structures within the emulator enables very accurate emulators for the WOPT and WWIT NPV constituents based on a relatively small number of simulations whilst also capturing the change in behaviour.

A comparison with Bayes linear emulation of the WOPT or WWIT NPV constituents, in each case fitted using all simulations, unlike the preliminary Bayes linear emulator in the structured emulation approach which is fitted using the green points in Figure 6 only, highlights the superior performance. Leave-one-out diagnostics are shown for the Bayes linear emulator of OLYMPUS 25 WOPTPROD2_2018_01, our running example, in Figure 8 depicting the emulator adjusted expectation with 95% (3 adjusted standard deviations) credible intervals versus the simulated output. The corresponding leave-one-out diagnostics plot using the structured emulation approach exploiting known simulator behaviour is shown in Figure 7a. Over the plateau region (green in 6b) the accuracy and credible interval width for the two

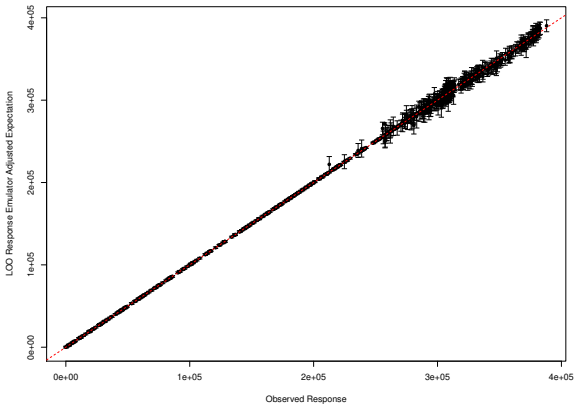


(a) Preliminary Bayes linear emulator predictive CI versus prod_2_2016_01.

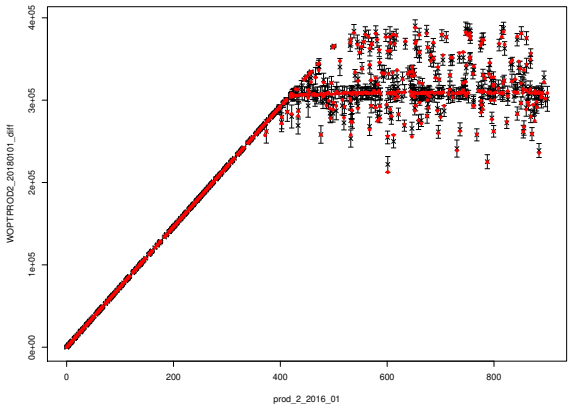


(b) Structured emulator with upper truncation predictive CI versus prod_2_2016_01.

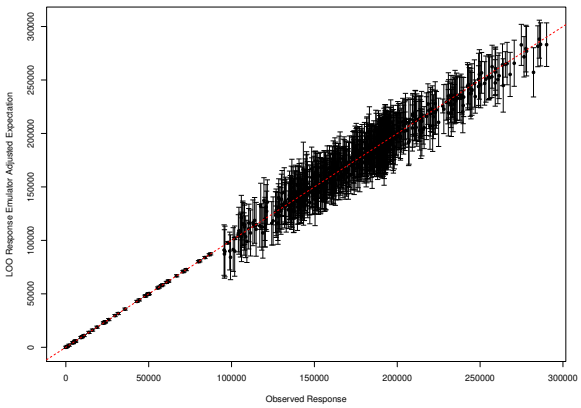
Figure 6: OLYMPUS 25 wave 1 WOPT for producer well 2 during the first two years (ending 01/01/2018) versus the corresponding target production rate, prod_2_2016_01. The top plot shows the wave 1 preliminary Bayes linear emulator predictive 3-sigma credible interval (CI) fitted using only the simulations with $d_{P2,2016} \geq c^u_{P2,2016}$. This is used within the structured emulation algorithm imposing the upper truncation due to the slope (black dotted line) with the CI shown in the bottom plot. The vertical blue and red lines are situated at $b_{P2,2016}$ and $c^u_{P2,2016}$ respectively. The purple, orange and green CI correspond to points in the slope, uncertain, and plateau regions respectively.



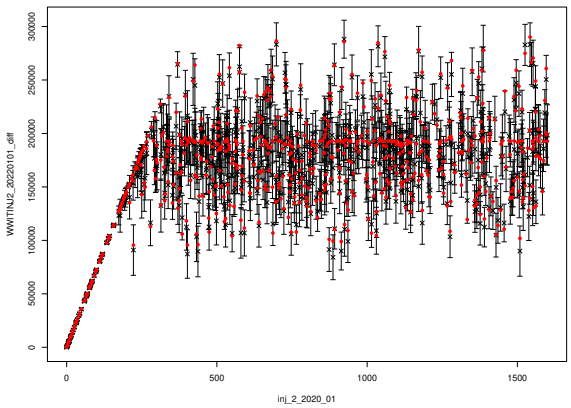
(a) WOPTPROD2_2018_01 emulator CI versus simulated output.



(b) WOPTPROD2_2018_01 emulator CI versus prod_2_2016_01.



(c) WWITINJ2_2022_01 emulator CI versus simulated output.



(d) WWITINJ2_2022_01 emulator CI versus inj_2_2020_01.

Figure 7: Structured emulation leave-one-out diagnostic plots for OLYMPUS 25 WOPTPROD2_2018_01 (top) and WWITINJ2_2022_01 (bottom). Left: Adjusted expectation with 95% credible intervals (CI) of width 3 adjusted standard deviations versus the simulated value. The red dashed line denotes equality of the emulator and simulator. Right: Credible interval versus the output's corresponding target production and injection rate respectively. Red points denote the simulated output.

emulators is comparable. However, within the slope (purple) and intermediate (orange) regions the Bayes linear emulator credible interval width is much wider. In the slope region there is at least a two orders of magnitude difference, a consequence of not imposing the known physical constructs. Moreover, within these two regions of parameter space the emulator adjusted

	WOPT	WWIT
OLYMPUS 25	$\leq 4.0\%$	$\leq 3.4\%$
OLYMPUS 33	$\leq 2.0\%^1$	$\leq 2.7\%$
OLYMPUS 45	$\leq 1.2\%^2$	$\leq 3.0\%^3$

Table 1: Summary of the maximum percentage of structured emulator with upper truncation 95% credible intervals which do not contain the simulated values in leave-one-out diagnostics for the WOPT and WWIT over the 8 control intervals for each of the 3 OLYMPUS models. The exceptions are: (1) for OLYMPUS 33 WOPT emulation of the output in one control interval yields a failure rate of 7.5%; (2) for OLYMPUS 45 WOPT emulation in two control intervals yields a failure rate of 6.1% & 5.3%; and (3) for OLYMPUS 45 WWIT emulation in one control interval yields a failure rate of 6.3%.

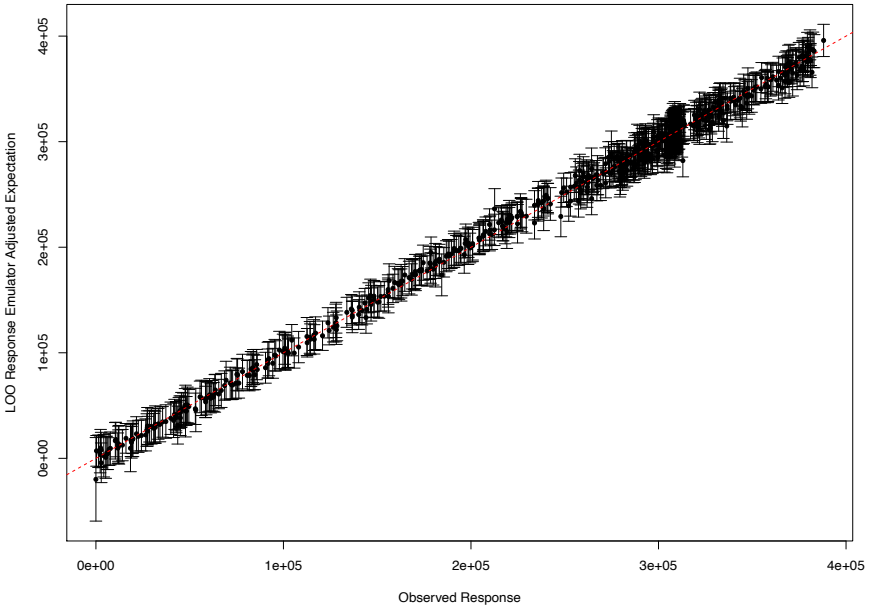


Figure 8: Bayes linear emulation leave-one-out diagnostic plot for OLYMPUS 25 WOPT-PROD2_2018_01 showing the adjusted expectation with 95% credible intervals of width 3 adjusted standard deviations versus the simulated output. The red dashed line denotes equality of the emulator and simulator. For comparison, the structured emulation exploiting known simulator behaviour leave-one-out diagnostics plot for the same output is shown in Figure 7a.

735 expectation often exceeds the maximum upper bound imposed by the target rate decision
736 parameter governing this period, whilst the majority of the credible interval upper bounds
737 also exceed this limit. This implies that unphysical emulator predictions are permitted which
738 may go unchecked. The structured emulation approach protects against this facet.

Structured behaviour is not observed for WWPT within a control interval since there is no corresponding target rate; its behaviour is a consequence of attempting to achieve a given target production rate subject to BHP constraints with water present within the oil field. We separately employ Bayes linear emulators for each of the WWPT constituents following the same approach as above, but fitted using all simulations in $\mathcal{D}$. For each OLYMPUS model the collection of 48 emulators for the WOPT, WWPT, and WWIT for each of the 8 control intervals and for wells in the CWG are combined following the "divide-and-conquer" approach in subsection 6.2.

8. Emulating Sums of Time Series Outputs.

8.1. Methodology. The "divide-and-conquer" approach in section 6 permits the exploitation of known behaviour such as illustrated in section 7. We develop accurate and efficient emulation methodology where the quantity of interest is the sum of time series outputs addressing the challenges arising from the discretisation of continuous time outputs. In subsection 8.1.1 we first emulate an approximation to the sum of time series outputs computed over longer time periods, thus reducing the number of emulators required, focusing on the merger of discounting intervals, before linking to the exact quantity of interest in subsection 8.1.2.

8.1.1. Emulation of an Average Discounting Approximation to the Sum of Time Series Outputs. Let $f(\mathbf{d})$ be the sum of time series outputs:

$$f(\mathbf{d}) = \sum_{i=1}^{N_t} \frac{1}{(1+d)^{\frac{t_i}{\tau}}} f_i(\mathbf{d}) \tag{8.1}$$

where i indexes the time point with $t_i < t_{i+1}$, N_t is the total number of discounting intervals, d is the discount factor, and τ the discounting period. This is analogous to in (6.1) following the "divide-and-conquer" approach with $q \equiv N_t$ and $a_i = (1+d)^{-\frac{t_i}{\tau}}$.

In situations where N_t is very large it may be impractical to accurately emulate and validate for all $f_i(\mathbf{d})$. An average discounting approximation to the exact quantity of interest, $f(\mathbf{d})$, is denoted by $\widetilde{f}(\mathbf{d})$ which is computed as a sum of outputs formed by amalgamating multiple time consecutive discounting intervals labelled by $\widetilde{f}_i(\mathbf{d})$. A formula for $\widetilde{f}(\mathbf{d})$ is given in (8.2), where $\tilde{N}_t < N_t$ is the numbered of combined time intervals, and λ_i is a weighted average discounting factor for the i^{th} interval defined in (8.3) for which k indexes the discounting intervals contained within the longer control interval, N_{t_i} is the total number of such discounting intervals, and $t_{i,0} = t_{i-1}$.

$$\widetilde{f}(\mathbf{d}) = \sum_{i=1}^{\tilde{N}_t} \lambda_i \widetilde{f}_i(\mathbf{d}) \tag{8.2}$$

$$\lambda_i = \frac{1}{t_i - t_{i-1}} \sum_{k=1}^{N_{t_i}} \frac{t_{i,k} - t_{i,k-1}}{(1+d)^{\frac{t_{i,k}}{\tau}}} \tag{8.3}$$

Note that using an averaged discount factor yields a more accurate approximation compared with applying the discounting at the end of each time interval. Assuming the $\widetilde{f}_i(\mathbf{d})$ are uncorrelated and using a collection of univariate emulators the adjusted expectation and variance formulae for $\widetilde{f}(\mathbf{d})$ are obtained following (6.2) and (6.3) respectively.

8.1.2. Linking the Exact and Approximate Sums of Time Series Outputs. The next step is to link $\widetilde{f}(\mathbf{d})$ with $f(\mathbf{d})$. Due to the similar form of (8.1) and (8.2) there exists a strong linear relationship between the approximate and exact $f(\mathbf{d})$ for which a simple linear regression in (8.4) provides a meaningful statistical link whilst capturing the additional induced uncertainties.

$$f(\mathbf{d}) = \beta_{0,\widetilde{f}} + \beta_{1,\widetilde{f}}\widetilde{f}(\mathbf{d}) + \varepsilon_{\widetilde{f}} \tag{8.4}$$

The adjusted expectation and variance are then computed using (8.5) and (8.6) respectively.

$$\mathbb{E}_{\mathbf{F}}[f(\mathbf{d})] = \hat{\beta}_{0,\widetilde{f}} + \hat{\beta}_{1,\widetilde{f}}\,\mathbb{E}_{\mathbf{F}}\left[\widetilde{f}(\mathbf{d})\right] \tag{8.5}$$

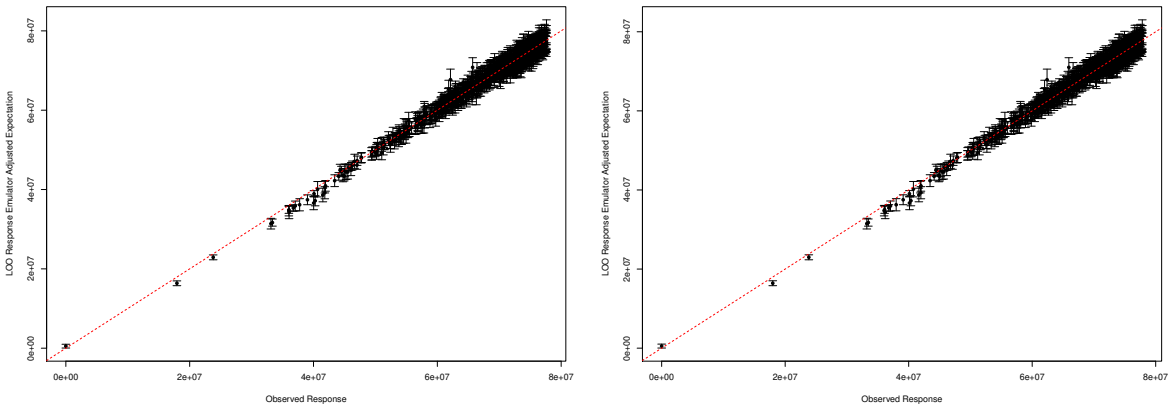
$$\begin{aligned}\mathrm{Var}_{\mathbf{F}}[f(\mathbf{d})] = {} & \mathrm{Var}\left[\hat{\beta}_{0,\widetilde{f}}\right] + 2\,\mathrm{Cov}\left[\hat{\beta}_{0,\widetilde{f}}, \hat{\beta}_{1,\widetilde{f}}\right]\mathbb{E}_{\mathbf{F}}\left[\widetilde{f}(\mathbf{d})\right] \\ & + \left\{\hat{\beta}^2_{1,\widetilde{f}} + \mathrm{Var}\left[\hat{\beta}_{1,\widetilde{f}}\right]\right\}\mathrm{Var}_{\mathbf{F}}\left[\widetilde{f}(\mathbf{d})\right] \\ & + \mathrm{Var}\left[\hat{\beta}_{1,\widetilde{f}}\right]\left(\mathbb{E}_{\mathbf{F}}\left[\widetilde{f}(\mathbf{d})\right]\right)^2 + \sigma^2_{\widetilde{f}}\end{aligned} \tag{8.6}$$

Estimates of the regression coefficients, $\hat{\beta}_{0,\widetilde{f}}$ and $\hat{\beta}_{1,\widetilde{f}}$, along with their variances and covariance, are obtained using the wave 0 exploratory simulations data, whilst $\varepsilon_{\widetilde{f}}$ is treated as independent with residual standard error $\sigma_{\widetilde{f}}$. The collection of all simulation data, $\mathbf{F}$, is as defined in subsection 6.1.

8.2. Results. The NPV objective function in the TNO OLYMPUS Well Control Optimisation Challenge (see (2.1) and (2.2)) is of the form of (8.1). For each OLYMPUS ensemble member $f(\mathbf{d}) = \mathrm{NPV}_j(\mathbf{d})$ and $f_i(\mathbf{d})$ are the NPV constituents. In this application the 8 decisions for each well are enacted over periods constructed by amalgamating consecutive 3-month discounting intervals. The by model average discounting approximate NPV, $\widetilde{\mathrm{NPV}}_j(\mathbf{d})$ is obtained from (6.4), noting that each of $f^{op}_{Pk,t_i}(\mathbf{d})$, $f^{wp}_{Ik,t_i}(\mathbf{d})$, and $f^{wi}_{Ik,t_i}(\mathbf{d})$ are calculated over periods longer than the discounting intervals, hence $f(\mathbf{d})$ does correspond to the approximation $\widetilde{f}(\mathbf{d})$, and with $a_i = \lambda_i$ from (8.3).

Emulation of $\widetilde{\mathrm{NPV}}_j(\mathbf{d})$ is performed following the method described in subsection 8.1 summing structured emulators for the WOPT and WWIT contributors (details in subsection 7.2 and Bayes linear emulators for the WWPT constituents. Leave-one-out diagnostics plots for the OLYMPUS 25 approximate NPV is shown in Figure 9a. There exists a strong linear relation between the emulator adjusted expectation and the simulated approximate NPV with the majority of points situated close to the red dashed equality line. It is observed that the uncertainty generally increases with the value of the approximate NPV. Petroleum reservoir engineering provides insight: higher target production and injection rates are generally necessary to achieve the largest NPVs. For the WOPT and WWIT structured emulators this occurs above the extrapolation cut-off and thus each constituent emulator exhibits a larger uncertainty. Furthermore, when many of the NPV constituents fall in their slope regions the structured emulator returns a small uncertainty determined by the tolerance. These linearly combine to produce a small uncertainty for the approximate NPV.

The exact and average discounting approximate NPV for each OLYMPUS model are linked using the simple linear regression framework in (8.4) where the coefficients are estimated



(a) Approx NPV CI versus simulated values.

(b) Exact NPV CI versus simulated values.

Figure 9: Emulator leave-one-out diagnostic plots for OLYMPUS 25 average discounting approximate NPV (left) and exact NPV (right, obtained via a linear model of the form in (8.4) on the emulated approximate NPV) showing the adjusted expectation with 3 adjusted standard deviation Credible Intervals (CI) versus simulated values. The red dashed line denotes when the emulator and simulator coincide.

	Approximate NPV	Exact NPV
OLYMPUS 25	6.7%	6.7%
OLYMPUS 33	4.4%	4.3%
OLYMPUS 45	3.7%	3.8%

Table 2: Summary of the percentage of emulator 95% credible intervals which do not contain the simulated values in leave-one-out diagnostics for the average discounting approximation to the NPV and the exact NPV for each of the 3 OLYMPUS models.

using the wave 0 simulation data. This accounts for the discrepancy induced by coalescing the discounting intervals. Leave-one-out emulator diagnostics for the OLYMPUS 25 NPV are shown in Figure 9b. The results are very similar to those for the approximate NPV with our commentary and interpretation mirroring the above. The percentage of 95% credible intervals containing the simulated value (computed using simulation output in the respective average discounting approximate or exact NPV formula) for each of the 3 OLYMPUS models is reported in Table 2.

9. Emulation of a Multi-Model Ensemble Mean.

9.1. Methodology. The objective is to emulate the ensemble mean output, $\bar{f}(\mathbf{d})$, by combining emulators for the individual models' outputs, $f_j(\mathbf{d})$. A reasonable assumption is that the ensemble members are independent given the complexity of their differing constructions.

824 **9.1.1. When Simulations are Available for all Ensemble Members.** When simulations
825 are relatively quick to evaluate; large amounts of computing resources are available, or there
826 is a desire to minimise the uncertainty (such as in the TNO OLYMPUS Challenge due to the
827 underlying geology, which is particularly relevant when the ensemble mean NPV is assumed
828 equal to the expected NPV), it may be possible to simulate from the entire ensemble. The
829 ensemble mean output is computed as either the arithmetic or a weighted mean (with weights
830 obtained from a prior probability distribution over the models) of the individual model out-
831 puts. This presents a natural method to emulate $\bar{f}(\mathbf{d})$ with the adjusted expectation and
832 variance defined in (9.1) and (9.2), where $\mathbf{F} = \{\mathbf{F}_j\}_{j=1}^N$ denotes all necessary simulation data
833 with $\mathbf{F}_j$ being the outputs for ensemble member j, and weights ω_j, with $\omega_j = \frac{1}{N}$ for the
834 arithmetic mean.

835 (9.1) $$\mathbb{E}_{\mathbf{F}}\left[\bar{f}(\mathbf{d})\right] = \sum_{j=1}^{N} \omega_j \, \mathbb{E}_{\mathbf{F}_j}[f_j(\mathbf{d})]$$

836 (9.2) $$\mathrm{Var}_{\mathbf{F}}\left[\bar{f}(\mathbf{d})\right] = \sum_{j=1}^{N} \omega_j^2 \, \mathrm{Var}_{\mathbf{F}_j}[f_j(\mathbf{d})]$$

837 The variance formula may be adapted when the output for different ensemble members are
838 believed to be correlated by introducing the relevant covariance terms in (9.2).

839 **9.1.2. Using an Ensemble Subsampling Linear Model.** A more realistic and practical
840 scenario is that simulations are only performed for a subset of the ensemble, such as, but not
841 limited to, those selected using the techniques described in subsection 4.1. A linear model
842 of the form shown in (4.1) is used to emulate $\bar{f}(\mathbf{d})$ with the emulated output for each of
843 the sub-selected models as inputs. These are $\{f_{j_1}(\mathbf{d}), \ldots, f_{j_{\tilde{N}}}(\mathbf{d})\}$, for which $\tilde{N} < N$, with
844 $j_1, \ldots, j_{\tilde{N}} \in \{1, \ldots, N\}$, and $j_k \neq j_l$ for $k \neq l$. The estimated coefficients are denoted by $\hat{\alpha}_{\mathrm{ES}}$
845 and $\hat{\beta}_{k,\mathrm{ES}}$. It is assumed that the individual emulator outputs and the regression coefficients
846 are uncorrelated, which is justifiable if two distinct simulation data sets are used to construct
847 the linear model and fit the emulators. Under this formulation the adjusted expectation is
848 shown in (9.3).

849 $$\begin{aligned}\mathbb{E}_{\mathbf{F}}\left[\bar{f}(\mathbf{d})\right] &= \mathbb{E}_{\mathbf{F}}\left[\hat{\alpha}_{\mathrm{ES}} + \sum_{k=1}^{\tilde{N}} \hat{\beta}_{k,\mathrm{ES}} f_{j_k}(\mathbf{d}) + \varepsilon_{\mathrm{ES}}(\mathbf{d})\right] \\ 850 \quad (9.3) \qquad &= \hat{\alpha}_{\mathrm{ES}} + \sum_{k=1}^{\tilde{N}} \hat{\beta}_{k,\mathrm{ES}} \, \mathbb{E}_{\mathbf{F}_{j_k}}[f_{j_k}(\mathbf{d})]\end{aligned}$$

851 Define $\hat{\boldsymbol{\beta}}_{\mathrm{ES}} = (\hat{\alpha}_{\mathrm{ES}}, \hat{\beta}_{1,\mathrm{ES}}, \ldots, \hat{\beta}_{\tilde{N},\mathrm{ES}})^{\mathrm{T}} \in \mathbb{R}^{\tilde{N}+1}$ with $\Sigma_{\beta,\mathrm{ES}} = \mathrm{Var}\left[\hat{\boldsymbol{\beta}}_{\mathrm{ES}}\right]$, and
852 $X_{\mathrm{ES}}(\mathbf{d}) = (1, f_{j_1}(\mathbf{d}), \ldots, f_{j_{\tilde{N}}}(\mathbf{d}))^{\mathrm{T}} \in \mathbb{R}^{\tilde{N}+1}$ with uncorrelated components, so $\mathrm{Var}_{\mathbf{F}}[X_{\mathrm{ES}}(\mathbf{d})]$
853 is diagonal. The adjusted variance is presented in (9.4), where $\hat{\sigma}_{\mathrm{ES}}$ is the estimated residual

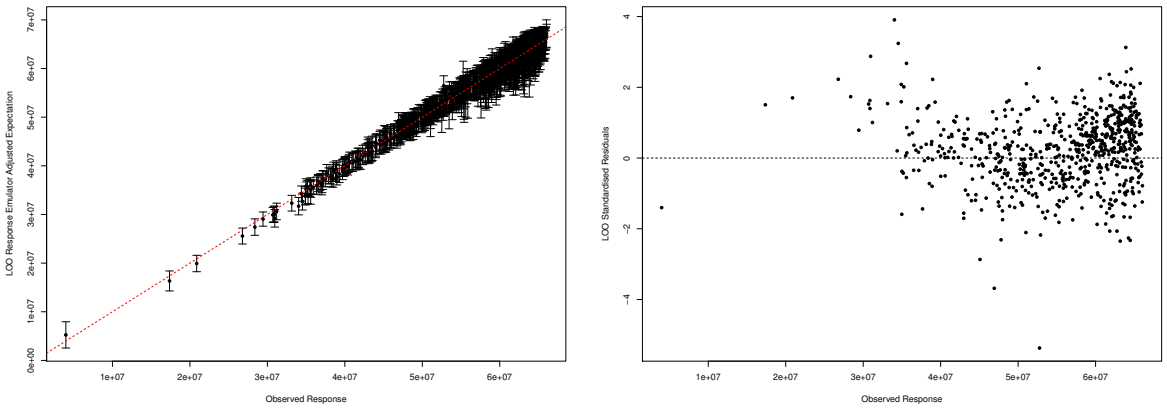
standard error for $\varepsilon_{\text{ES}}(\mathbf{d})$.

$$
\begin{aligned}
\text{Var}_{\mathbf{F}}\left[\bar{f}(\mathbf{d})\right] = \text{Var}\left[\hat{\alpha}_{\text{ES}}\right] & \\
&+ \sum_{k=1}^{\tilde{N}} \text{Var}\left[\hat{\beta}_{k,\text{ES}}\right] \left(\text{Var}_{\mathbf{F}_{j_k}}\left[f_{j_k}(\mathbf{d})\right] + \mathbb{E}_{\mathbf{F}_{j_k}}\left[f_{j_k}(\mathbf{d})\right]^2\right) \\
&+ 2\sum_{k=1}^{\tilde{N}} \text{Cov}\left[\hat{\alpha}_{\text{ES}}, \hat{\beta}_{k,\text{ES}}\right] \mathbb{E}_{\mathbf{F}_{j_k}}\left[f_{j_k}(\mathbf{d})\right] \\
&+ \sum_{\substack{k \neq l \\ k,l=1,\ldots,\tilde{N}}} \left(\text{Cov}\left[\hat{\beta}_{k,\text{ES}}, \hat{\beta}_{l,\text{ES}}\right] \mathbb{E}_{\mathbf{F}_{j_k}}\left[f_{j_k}(\mathbf{d})\right] \mathbb{E}_{\mathbf{F}_{j_l}}\left[f_{j_l}(\mathbf{d})\right]\right) \\
&+ \sum_{k=1}^{\tilde{N}} \hat{\beta}_k^2 \, \text{Var}_{\mathbf{F}_{j_k}}\left[f_{j_k}(\mathbf{d})\right] + \hat{\sigma}_{\text{ES}}^2
\end{aligned}
\tag{9.4}
$$

9.2. Results. The process of building structured emulators for each of the NPV constituents in subsection 7.2, their combination via the NPV formula to obtain the average discounting approximate NPV, and subsequent linking to the exact NPV in subsection 8.2, is repeated for each of the three sub-selected OLYMPUS models. For the TNO OLYMPUS Well Control Optimisation Challenge and our decision support setup the ensemble mean NPV, $\bar{f}(\mathbf{d}) = \overline{\text{NPV}}(\mathbf{d})$, is the quantity of interest as the objective and utility function respectively. This is emulated using the ensemble subsampling linear model devised in subsection 4.2 to combine the emulators for the OLYMPUS 25, 33, & 45 NPVs, following the approach in subsection 9.1.

It is not possible to perform leave-one-out diagnostics for the true ensemble mean NPV because simulations have only been performed for the identified subset of OLYMPUS models. The wave 0 simulations were run for all 50 OLYMPUS models under a setup using a shorter field lifetime due to available computational resources, hence these cannot be used in emulator diagnostics. Note that the additional uncertainty pertaining to the ensemble subsampling linear model is accounted for within the hierarchical emulator construction. Instead we compare the hierarchical emulator with the predicted ensemble mean NPV in Figure 10 where Figure 10a demonstrates accurate predictions. Moreover, the increase in the uncertainty compared to individually emulating a single OLYMPUS model NPV, such as for OLYMPUS 25 NPV in Figure 9b, is modest; thus the process of subsampling from the ensemble before reconstructing the ensemble mean NPV contributes relatively little additional uncertainty versus the structured emulation of the NPV constituents for each model. Figure 10b shows no distinguishable pattern in the pseudo standardised residuals, whilst the majority are of magnitude less than three.

9.3. Emulator Comparison. Two approaches were implemented for emulating the ensemble mean NPV: a Bayes linear emulator in subsection 3.2; and a hierarchical emulator which exploits known constrained behaviour for certain simulator outputs built up over subsections 4.2, 6.2, 7.2, 8.2, and 9.2. Firstly, comparing each emulator's adjusted variances evaluated for the same large collection of decision parameter vectors in Figure 11 demonstrates



(a) Hierarchical emulator CI versus subsampling predicted ensemble mean NPV. The red dashed line denotes emulator and simulator equality.

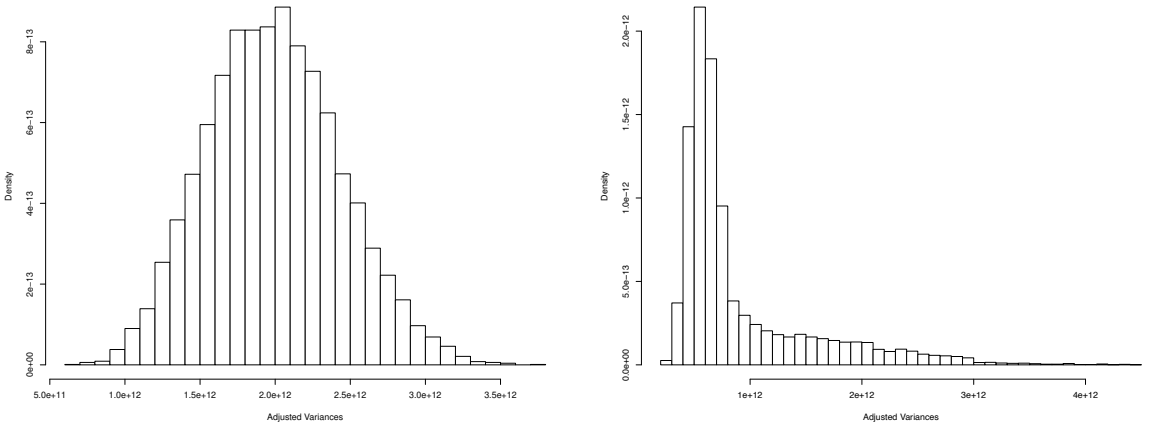
(b) Hierarchical emulator standardised residuals versus the simulated ensemble mean NPV.

Figure 10: OLYMPUS wave 1 hierarchical emulation diagnostics plots for the predicted ensemble mean NPV via the ensemble subsampling linear model combining the emulation output for the exact NPV of the three sub-selected OLYMPUS models.

how the hierarchical emulator achieves a discernible reduction in the uncertainty versus the Bayes linear emulator. This feature is also evident when comparing the leave-one-out diagnostics plots in Figures 2 and 10a where there is a prevalent reduction in the credible interval widths. A direct comparison of the adjusted variances for each decision parameter vector highlights an average reduction in the adjusted variance of more than a half. Note that there exist a small number of cases where there is a moderate increase in the uncertainty, although this is outweighed by the gains achieved across the majority of sampled locations within the decision parameter space.

A crucial motivation for employing emulators as a surrogate to computer models is their speed of evaluation in order to enable further analyses such as decision support. Bayes linear emulation is known to be a very fast and efficient means of constructing emulators. In this application we achieve a substantial reduction in computation time with over 2000 emulator evaluations for new decision parameter settings per second using a single core of a standard desktop computer or laptop. This is juxtaposed with approximately 30 minutes per OLYMPUS model simulation, or 25 hours when using the entire ensemble. The combination of ensemble subsampling and Bayes linear emulation equates to an efficiency gain of the order of 10^8.

The full hierarchical emulation process applied to the ensemble mean NPV requires for each OLYMPUS model the fitting of 48 separate emulators: 32 of the structured type; and 16 Bayes linear emulators, a total of 144 emulators over the three sub-selected OLYMPUS models. Next, these are combined to obtain emulators for the approximate and exact NPVs



(a) Bayes linear emulator.

(b) Hierarchical emulator.

Figure 11: Histograms comparing the Bayes linear and hierarchical emulators adjusted variances for the ensemble mean NPV. Note that the seemingly large variances are inline with the simulated ensemble mean NPV which is of the order 5.0×10^7 \$ to 6×10^7 \$.

for each OLYMPUS model, before emulating the ensemble mean NPV, and then linking to the expected NPV. The computational performance is more modest achieving emulator evaluations at approximately 4 new decision parameter vectors per second using a single core, and is thus slower than Bayes linear emulation. However, in comparison with direct simulation from the OLYMPUS ensemble there is a considerable efficiency improvement of the order of 10^4. This is sufficient for comprehensively exploring the decision parameter space. Moreover, the additional computational expense of hierarchical emulation can be justified by the reduction in emulator uncertainty. This is highly beneficial to performing an iterative decision support analysis where reducing emulator uncertainty is imperative for efficiently eliminating non-implausible regions of the decision parameter space, thus avoiding extra waves of extremely expensive simulations at locations that would have been ruled out by more accurate emulators. The additional computational cost is therefore offset versus the need for extra simulations. Such arguments are also relevant to analyses using single-stage (or one-shot) designs where fewer expensive computer model evaluations are required to achieve similar emulator accuracy across the parameter space. Both forms of emulators are easily parallelisable, thus permitting further efficiency gains.

10. Conclusion. We have presented a methodological toolkit for the analysis of multi-model ensembles of "grey-box" computer models. This include: an efficient technique for obtaining a small representative subset of models by subsampling from a multi-model ensemble; targeted Bayesian design methodology incorporating relevant prior information to the objective of providing decision support under uncertainty; a "divide-and-conquer" approach to emulation of sums of outputs where it is preferable to emulate the constituents, for example due to knowledge of their underlying behaviour; structured emulation of outputs to

exploit constrained and structured behaviour through the partitioning of the parameter space and use of truncated emulators; the efficient combination of multiple emulators for time series outputs through an average discounting approximation; and emulation of an ensemble mean output. Combining these methods yields a novel hierarchical emulator achieving more accurate predictions for quantities of interest, whilst each technique may also be employed separately depending on the problem specific features exhibited by the computer model.

This is motivated by and applied to the TNO OLYMPUS Well Control Optimisation Challenge from the petroleum industry where the aim is to maximise the expected NPV, approximated by the ensemble mean NPV, as a function of well control decision parameters. We reconstrue this as a decision support problem where the utility function consists of a discounted sum of oil production, water injection, and water production, both by well and control interval. The first two simulator outputs exhibit partially known behaviour, constrained by choices of inputs and physical limits with respect to their corresponding target production and injection rate decision parameters respectively, with this feature encompassed within our structured emulator formulation. The application demonstrates superior accuracy versus Bayes linear emulators, whilst the slower speed of evaluation is mitigated by the need for fewer (waves of) simulations from the expensive computer model ensemble. Both factors are important for the overall aim of providing robust decision support under uncertainty. Moreover, we introduce multi-model ensemble subsampling techniques to efficiently identify a representative subset of models which collectively best characterises the ensemble mean output of interest, in this application, the ensemble mean NPV, whilst also providing a method for their prediction. This constitutes a novel application to the petroleum industry where multi-model ensembles are commonly used to represent geological uncertainty, greatly reducing the computational cost of our decision analysis.

The next step is to employ the presented hierarchical emulation methodology within iterative decision support, applied to the TNO OLYMPUS Well Control Optimisation Challenge, which incorporates a comprehensive and realistic uncertainty quantification to statistically link inferences for the computer model (OLYMPUS) with the corresponding real world physical system. See [40, Sec. 4.6 & 4.7] for details. In addition, further methodological development should focus on enhancing the overall hierarchical emulation framework. This may be achieved by revising the structured emulators change point estimation methods, classification and truncation, as well as via the refinement of the uncertainty propagation in subsections 8.1 and 9.1. Another direction is multivariate structured emulation of the NPV constituents to assess their correlation and thus more accurately quantify the approximate NPV by ensemble member emulator variance in subsections 7.2 and 8.2. Such further methodological developments must also be efficient so as not add to the computational burden.

The methodological toolkit and their combination to form a hierarchical emulator presented in this paper, whilst motivated by and tailored to the petroleum well control optimisation problem, is sufficiently flexible and adaptable to handle other (partially) known forms of computer model outputs and functions thereof. Opening "black-box" simulators and exploring functions of their output to investigate their behaviours is evidently beneficial, as is using domain expert prior knowledge and small carefully designed collections of simulations. Another example of emulating "grey-box" models is in known boundary emulation [58, 29]. The additional prior information can then be used to guide the choice from existing emulation

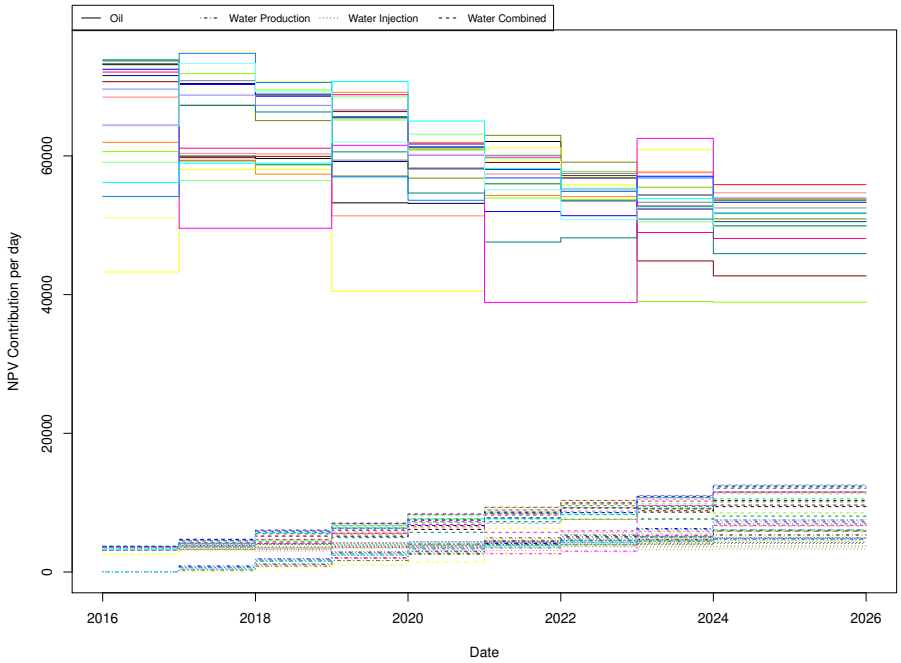
methods or to design novel forms which exploit known behavioural facets in order to achieve superior accuracy and enhance the usefulness of emulators for real world applications.

Appendix A. TNO OLYMPUS Well Control Optimisation Challenge – Extended Results.

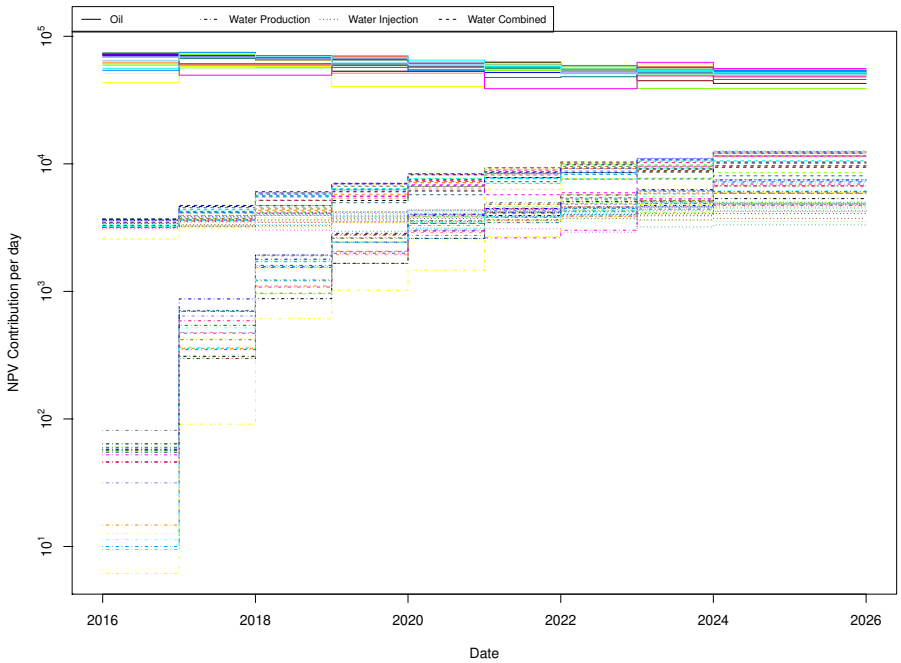
In this appendix we extend our discussion of results of the application to the TNO OLYMPUS Well Control Optimisation Challenge (see section 2 for an overview) using the methodology proposed in this paper.

A.1. OLYMPUS Exploratory Analysis – Additional Plots and Discussion. Our exploratory analysis identifies large differences in the absolute contributions of oil and water, both production and injection, to the NPV objective function. This feature has potential ramifications for emulation and decision support. An assessment of the absolute contributions approximated within one year intervals for the OLYMPUS 25 NPV is shown in Figure 12 using each of the 20 exploratory analysis decision parameter vectors represented by different colours. In (2.3) the oil contribution (solid lines), $|Q_{j,op}(\mathbf{d}, t_i) \cdot r_{op}|$, is dominant versus both the absolute water production contribution (dot-dashed lines), $|Q_{j,wp}(\mathbf{d}, t_i) \cdot r_{wp}|$, and water injection contribution (dotted lines), $|Q_{j,wi}(\mathbf{d}, t_i) \cdot r_{wi}|$, as well as their sum (dashed lines), $|Q_{j,wp}(\mathbf{d}, t_i) \cdot r_{wp} + Q_{j,wi}(\mathbf{d}, t_i) \cdot r_{wi}|$. For earlier time intervals the magnitude of the oil contribution to the NPV is typically of the order of 100 times the combined water contribution which decays towards 10 times larger for later time intervals. Plotting on the logarithmic scale in Figure 12b facilitates an easier comparison of the water contributions. It is observed that water injection contributes a much larger amount to the NPV, particularly for earlier time intervals. This is to be expected since production wells are drilled within regions containing a high oil concentration, hence at initial times there should be very little water production. At later times the contribution becomes more alike as an increased quantity of water is produced in order to maintain oil production, whilst also noting the higher fixed cost per barrel of water produced versus injected. Similar observations are made for other OLYMPUS ensemble members.

A.2. Subsampling from Geological Multi-Model Ensembles – Additional Plots. Preliminary graphical investigations utilise plots of the ensemble mean versus the individual model over a range of outputs of interest for the wave 0 simulations. Examples of these plots are shown in Figure 13 where the black line denotes equality between the ensemble mean and individual ensemble member model output. The main outputs of interest stem from the NPV objective function and include: the ensemble mean NPV, oil production, water production and injection totals, both for the field and by well, as well as over the entire field lifetime, and for control intervals. Note that this is a preliminary graphical assessment which is limited to identifying one-dimensional relationships. Figures 13a to 13c show strong linear relationships with fairly limited variation providing evidence that even as individual models, OLYMPUS 25, 33 & 45 are potentially representative for the ensemble mean. An appropriate (linear) transformation may be applied in the cases seen in Figures 13b and 13c. In contrast OLYMPUS 50 does not appear to be a good representative model, at least individually, as seen in Figure 13d where the relationship is more challenging to model. This graphical investigation is also useful as a preliminary screening technique yielding a subset of 9 models to investigate

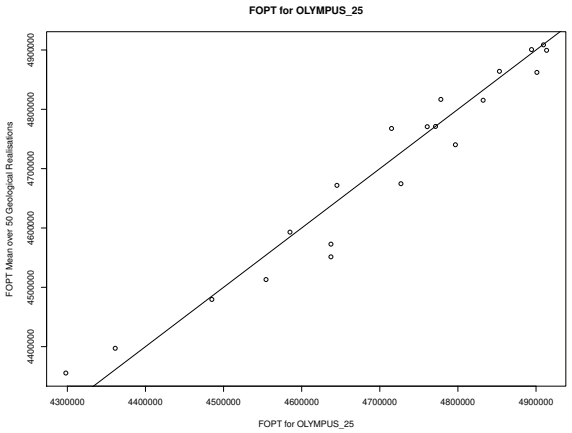


(a) Raw scale.

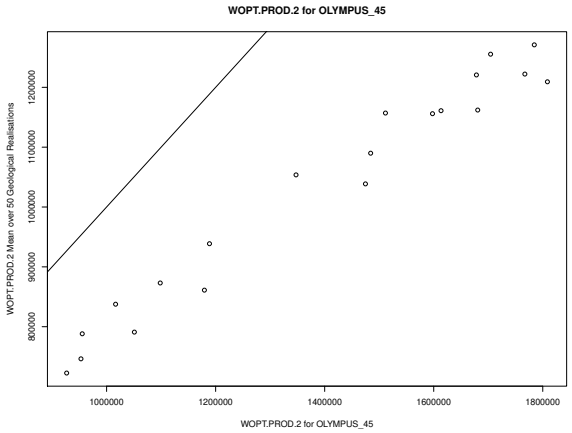


(b) Logarithmic scale.

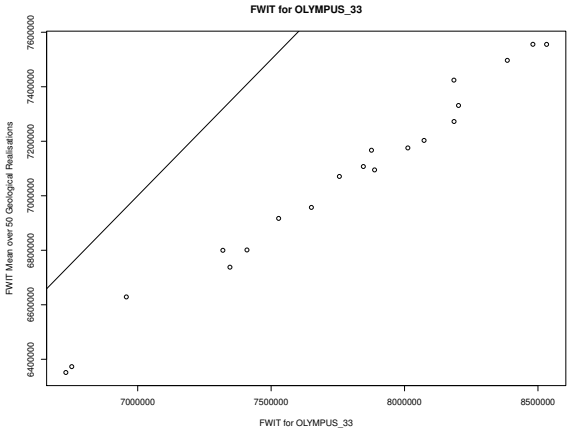
Figure 12: OLYMPUS 25 approximate absolute contribution to the NPV per year for each of the exploratory simulations shown as coloured lines. The NPV is decomposed into the oil production (solid line), absolute water production (dot-dashed line) and injection (dotted line), and the total water contribution (dashed line), with each scaled by the respective fixed NPV cost parameter. These are $|Q_{j,op}(\mathbf{d}, t_i) \cdot r_{op}|$, $|Q_{j,wp}(\mathbf{d}, t_i) \cdot r_{wp}|$, $|Q_{j,wi}(\mathbf{d}, t_i) \cdot r_{wi}|$ and $|Q_{j,wp}(\mathbf{d}, t_i) \cdot r_{wp} + Q_{j,wi}(\mathbf{d}, t_i) \cdot r_{wi}|$ in (2.3) respectively. The top and bottom plots are on the raw and logarithmic scale respectively.



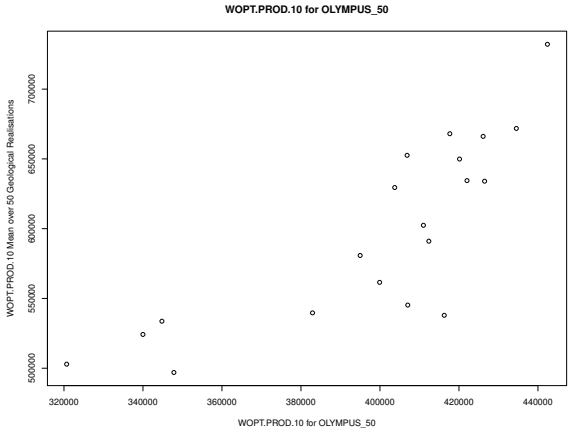
(a) OLYMPUS 25 FOPT.



(b) OLYMPUS 45 WOPTPROD2.



(c) OLYMPUS 33 FWIT.



(d) OLYMPUS 50 WOPTPROD10.

Figure 13: Subsampling from the multi-model OLYMPUS ensemble preliminary graphical investigations showing the ensemble mean versus individual model outputs. The black line denotes equality between the ensemble mean and individual model outputs. Note that in Figure 13d the black line is not shown due to the much smaller values of WOPTPROD10 for OLYMPUS 50 compared with the ensemble mean.

further: OLYMPUS 2, 6, 11, 23, 25, 33, 35, 37, & 38.

The combination of different OLYMPUS models is assessed using the linear model subsampling technique in 4.1. This is first applied to the above proposed subset of 9 OLYMPUS models before considering all models in a both directions stepwise selection with AIC. Only $\tilde{N} = 3$ models are necessary for a large number of the investigated outputs, as demonstrated in Figure 14 showing the linear model adjusted R^2 values for various outputs. All are high with

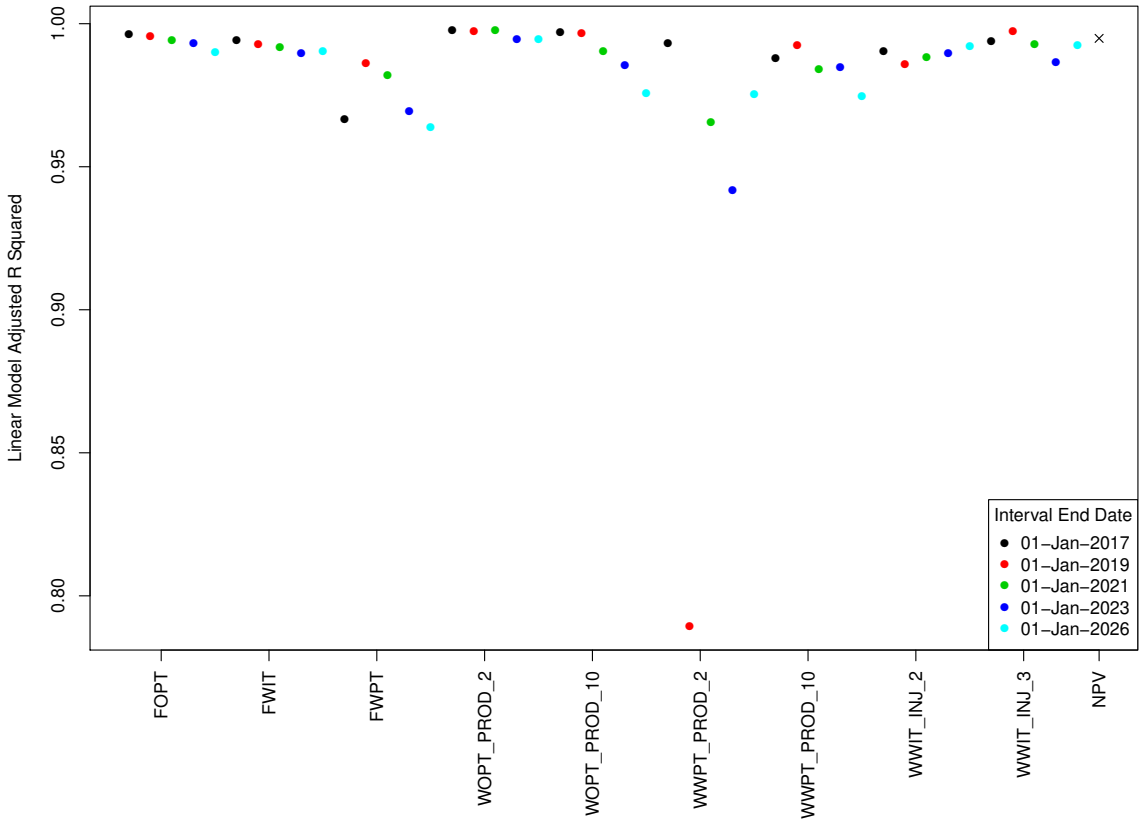


Figure 14: Adjusted R^2 values for the OLYMPUS ensemble subsampling linear models of the form in (4.1) for the ensemble mean of various outputs within control intervals using the same subset of $\tilde{N} = 3$ OLYMPUS models as predictors.

most greater than 0.95 implying that the majority of the ensemble variation can be explained by a small subset. These are OLYMPUS 25, 33, & 45. It is noted that OLYMPUS 25 & 33 were identified as part of the proposed subset of models, where as OLYMPUS 45 was not. This is because for certain outputs it was judged that OLYMPUS 45 did not provide a sufficiently good representation of the ensemble mean, however, in combination with OLYMPUS 25 & 33 via the linear models, these models collectively provide a good characterisation of the ensemble mean NPV, as well as other outputs.

A.3. Hierarchical Emulation of the Expected NPV – Additional Plots. The structured emulation technique incorporating known simulator behaviour in subsection 7.1 is applied separately for each of the OLYMPUS models to the WOPT and WWIT within each control interval for wells in the CWG, with simulator outputs considered as the $f(\mathbf{d})$. Firstly, conservative estimates for the change point upper bounds are calculated from the wave 1

simulations using (7.1), each time with $\delta_u = 10$. This ensures numerical stability and that an upper bound is obtained with all points exceeding this definitely in the plateau region. Next the extrapolation cut-offs are estimated as the change point lower bounds, via (7.2), with $\delta_l = 10$ to account for numerical precision within the simulations. The change point upper bounds and extrapolation cut-offs are illustrated for all WOPT and WWIT constituents for each wave 1 sub-selected OLYMPUS model in Figure 15 highlighting the region in which the "true" change point is believed to be situated.

REFERENCES

[1] Z. Alrashdi, M. Sayyafzadeh, and D. Guerillot, *Well control, field development and joint optimization using (μ+λ) evolutionary strategy algorithm and a stochastic rank*, European Association of Geoscientists and Engineers (EAGE) and the Netherlands Organisation for Applied Scientific Research (TNO), 2018, https://doi.org/10.3997/2214-4609.201802291, https://www.earthdoc.org/content/papers/10.3997/2214-4609.201802291.

[2] I. Andrianakis and P. G. Challenor, *The effect of the nugget on Gaussian process emulators of computer models*, Computational Statistics & Data Analysis, 56 (2012), pp. 4215–4228, https://doi.org/10.1016/j.csda.2012.04.020.

[3] I. Andrianakis, I. R. Vernon, N. McCreesh, T. J. McKinley, J. E. Oakley, R. N. Nsubuga, M. Goldstein, and R. G. White, *Bayesian History Matching of Complex Infectious Disease Models Using Emulation: A Tutorial and a Case Study on HIV in Uganda*, PLOS Computational Biology, 11 (2015), https://doi.org/10.1371/journal.pcbi.1003968.

[4] I. Andrianakis, I. R. Vernon, N. McCreesh, T. J. McKinley, J. E. Oakley, R. N. Nsubuga, M. Goldstein, and R. G. White, *History matching of a complex epidemiological model of human immunodeficiency virus transmission by using variance emulation*, Journal of the Royal Statistical Society: Series C (Applied Statistics), 66 (2017), pp. 717–740, https://doi.org/10.1111/rssc.12198.

[5] B. Ankenman, B. L. Nelson, and J. Staum, *Stochastic kriging for simulation metamodeling*, Operations Research, 58 (2010), pp. 371–382, https://doi.org/10.1287/opre.1090.0754, https://pubsonline.informs.org/doi/10.1287/opre.1090.0754.

[6] E. Baker, P. Barbillon, A. Fadikar, R. B. Gramacy, R. Herbei, D. Higdon, J. Huang, L. R. Johnson, P. Ma, A. Mondal, B. Pires, J. Sacks, and V. Sokolov, *Analyzing stochastic computer models: A review with opportunities*, Statistical Science, 37 (2022), pp. 64–89, https://doi.org/10.1214/21-STS822.

[7] L. S. Bastos and A. O'Hagan, *Diagnostics for Gaussian Process Emulators*, Technometrics, 51 (2009), pp. 425–438, https://doi.org/10.1198/TECH.2009.08019.

[8] M. Binois, R. B. Gramacy, and M. Ludkovski, *Practical heteroscedastic gaussian process modeling for large simulation experiments*, Journal of Computational and Graphical Statistics, 27 (2018), pp. 808–821, https://doi.org/10.1080/10618600.2018.1458625, https://www.tandfonline.com/doi/full/10.1080/10618600.2018.1458625.

[9] P. S. Craig, M. Goldstein, J. C. Rougier, and A. H. Seheult, *Bayesian Forecasting for Complex Systems Using Computer Simulators*, Journal of the American Statistical Association, 96 (2001), pp. 717–729, https://doi.org/10.1198/016214501753168370.

[10] P. S. Craig, M. Goldstein, A. H. Seheult, and J. A. Smith, *Bayes linear strategies for matching hydrocarbon reservoir history*, in Bayesian Statistics 5, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, eds., no. 5 in Proceedings of the Valencia International Meeting, Clarendon Press, 1996, pp. 69–95.

[11] P. S. Craig, M. Goldstein, A. H. Seheult, and J. A. Smith, *Pressure matching for hydrocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments*, in Case Studies in Bayesian Statistics, C. Gatsonis, J. S. Hodges, R. E. Klass, R. E. McCulloch, P. Rossi, and N. D. Singpurwalla, eds., vol. 121 of Lecture Notes in Statistics, Springer-Verlag, 1st ed., 1997, pp. 37–93, https://doi.org/10.1007/978-1-4612-2290-3.

[12] P. S. Craig, M. Goldstein, A. H. Seheult, and J. A. Smith, *Constructing partial prior specifications for models of complex physical systems*, Journal of the Royal Statistical Society: Series D (The Statistician), 47 (1998), pp. 37–53, https://doi.org/10.1111/1467-9884.00115.
[13] J. Cumming and M. Goldstein, *Bayes Linear Uncertainty Analysis for Oil Reservoirs Based on Multisale Computer Experiments*, in The Oxford Handbook of Applied Bayesian Analysis, A. O'Hagan and M. West, eds., Oxford University Press, 2010, ch. 10, pp. 241–270.
[14] T. L. Edwards, M. A. Brandon, G. Durand, N. R. Edwards, N. R. Golledge, P. B. Holden, O. J. Nias, A. J. Payne, C. Ritz, and A. Wernecke, *Revisiting Antarctic ice loss due to marine ice-cliff instability*, Nature, 566 (2019), pp. 58–64, https://doi.org/10.1038/s41586-019-0901-4.
[15] European Association of Geoscientists and Engineers (EAGE) and the Netherlands Organisation for Applied Scientific Research (TNO), *EAGE/TNO Workshop on OLYMPUS Field Development Optimization*, EAGE Publications, 2018.
[16] B. D. Finetti, *Theory of Probability*, vol. 1, John Wiley & Sons Ltd, 1974.
[17] B. D. Finetti, *Theory of Probability*, vol. 2, John Wiley & Sons Ltd, 1975.
[18] M. Goldstein, *Bayes Linear Analysis*, in Encyclopedia of Statistical Sciences, S. Kotz, C. B. Read, N. Balakrishnan, B. Vidakovic, and N. L. Johnson, eds., vol. 3, Wiley, 2nd ed., 2006, pp. 29–34, https://doi.org/10.1002/0471667196.ess0986.pub2.
[19] M. Goldstein, *Subjective Bayesian Analysis: Principles and Practice*, Bayesian Analysis, 1 (2006), pp. 403–420, https://doi.org/10.1214/06-BA116.
[20] M. Goldstein, *External Bayesian analysis for computer simulators*, in Bayesian Statistics 9, J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, eds., no. 9 in Proceedings of the Valencia International Meeting, Oxford University Press, 2011, pp. 201–228, https://doi.org/10.1093/acprof:oso/9780199694587.001.0001.
[21] M. Goldstein and J. Rougier, *Bayes linear calibrated prediction for complex systems*, Journal of the American Statistical Association, 101 (2006), pp. 1132–1143, https://doi.org/10.1198/016214506000000203.
[22] M. Goldstein and D. Wooff, *Bayes Linear Statistics: Theory and Methods*, Wiley Series in Probability and Statistics, Wiley, 2007, https://doi.org/10.1002/9780470065662.
[23] R. B. Gramacy and H. K. H. Lee, *Bayesian Treed Gaussian Process Models with an Application to Computer Modeling*, Journal of the American Statistical Association, 103 (2008), pp. 1119–1130, https://doi.org/10.2307/27640148.
[24] R. B. Gramacy and H. K. H. Lee, *Cases for the nugget in modeling computer experiments*, Statistics and Computing, 22 (2012), pp. 713–722, https://doi.org/10.1007/s11222-010-9224-x.
[25] A. Harb, H. Kassem, and K. Ghorayeb, *Olympus field development optimization challenge – american university of beirut*, European Association of Geoscientists and Engineers (EAGE) and the Netherlands Organisation for Applied Scientific Research (TNO), 2018, https://doi.org/10.3997/2214-4609.201802292, https://www.earthdoc.org/content/papers/10.3997/2214-4609.201802292.
[26] T. Harris, B. Li, and R. Sriver, *Multimodel ensemble analysis with neural network gaussian processes*, The Annals of Applied Statistics, 17 (2023), pp. 3402–3425, https://doi.org/10.1214/23-AOAS1768, https://projecteuclid.org/journals/annals-of-applied-statistics/volume-17/issue-4/Multimodel-ensemble-analysis-with-neural-network-Gaussian-processes/10.1214/23-AOAS1768.full.
[27] K. Heitmann, D. Higdon, M. White, S. Habib, B. J. Williams, E. Lawrence, and C. Wagner, *The Coyote Universe II: Cosmological Models and Precision Emulation of the Nonlinear Matter Power Spectrum*, The Astrophysical Journal, 705 (2009), pp. 156–174, https://doi.org/10.1088/0004-637x/705/1/156.
[28] D. Higdon, J. Gattiker, B. Williams, and M. Rightley, *Computer Model Calibration Using High-Dimensional Output*, Journal of the American Statistical Association, 103 (2008), pp. 570–583, https://doi.org/10.1198/016214507000000888.
[29] S. E. Jackson and I. Vernon, *Efficient Emulation of Computer Models Utilising Multiple Known Boundaries of Differing Dimensions*, Bayesian Analysis, 18 (2023), pp. 165–191, https://doi.org/10.1214/22-BA1304.
[30] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*, vol. 1 of Wiley Series in Probability and Statistics, Wiley, 2nd ed., 1994.
[31] C. G. Kaufman, D. Bingham, S. Habib, K. Heitmann, and J. A. Frieman, *Efficient emulators of*

computer experiments using compactly supported correlation functions, with an application to cosmology, The Annals of Applied Statistics, 5 (2011), pp. 2470–2492, https://doi.org/10.1214/11-AOAS489.

[32] J. C. Kennedy, D. A. Henderson, and K. J. Wilson, *Multilevel emulation for stochastic computer models with application to large offshore wind farms*, Journal of the Royal Statistical Society: Series C (Applied Statistics), 72 (2023), pp. 608–627, https://doi.org/10.1093/jrsssc/qlad023.

[33] M. C. Kennedy and A. O'Hagan, *Bayesian calibration of computer models*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63 (2001), pp. 425–464, https://doi.org/10.1111/1467-9868.00294.

[34] T. L., S. Nowicki, B. Marzeion, R. Hock, H. Goelzer, H. Seroussi, N. C. Jourdain, D. A. Slater, F. E. Turner, C. J. Smith, C. M. McKenna, E. Simon, A. Abe-Ouchi, J. M. Gregory, E. Larour, W. H. Lipscomb, A. J. Payne, A. Shepherd, C. Agosta, P. Alexander, T. Albrecht, B. Anderson, X. Asay-Davis, A. Aschwanden, A. Barthel, A. Bliss, R. Calov, C. Chambers, N. Champollion, Y. Choi, R. Cullather, J. Cuzzone, C. Dumas, D. Felikson, X. Fettweis, K. Fujita, B. K. Galton-Fenzi, R. Gladstone, N. R. Golledge, R. Greve, T. Hattermann, M. J. Hoffman, A. Humbert, M. Huss, P. Huybrechts, W. Immerzeel, T. Kleiner, P. Kraaijenbrink, S. L. clec'h, V. Lee, G. R. Leguy, C. M. Little, D. P. Lowry, J.-H. Malles, D. F. Martin, F. Maussion, M. Morlighem, J. F. O'Neill, I. Nias, F. Pattyn, T. Pelle, S. F. Price, A. Quiquet, V. Radić, R. Reese, D. R. Rounce, M. Rückamp, A. Sakai, C. Shafer, N.-J. Schlegel, S. Shannon, R. S. Smith, F. Straneo, S. Sun, L. Tarasov, L. D. Trusel, J. V. Breedam, R. van de Wal, M. van den Broeke, R. Winkelmann, H. Zekollari, C. Zhao, T. Zhang, and T. Zwinger, *Projected land ice contributions to twenty-first-century sea level rise*, Nature, 593 (2021), pp. 74–82, https://doi.org/10.1038/s41586-021-03302-y, https://www.nature.com/articles/s41586-021-03302-y.

[35] J. E. Oakley, *Decision-Theoretic Sensitivity Analysis for Complex Computer Models*, Technometrics, 51 (2009), pp. 121–129, https://doi.org/10.1198/TECH.2009.0014.

[36] J. E. Oakley and A. O'Hagan, *Probabilistic sensitivity analysis of complex models: a Bayesian approach*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 66 (2004), pp. 751–769, https://doi.org/10.1111/j.1467-9868.2004.05304.x.

[37] A. O'Hagan, *Bayesian analysis of computer code outputs: A tutorial*, Reliability Engineering and System Safety, 91 (2006), pp. 1290–1300, https://doi.org/10.1016/j.ress.2005.11.025.

[38] A. O'Hagan, M. C. Kennedy, and J. E. Oakley, *Uncertainty Analysis and other Inference Tools for Complex Computer Codes (with discussion)*, in Bayesian Statistics 6, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, eds., no. 6 in Proceedings of the Valencia International Meeting, Clarendon Press, 1998-08-13, pp. 503–524.

[39] J. Onwunalu, *Stochastic oilfield optimization for hedging against uncertain future development plans*, European Association of Geoscientists and Engineers (EAGE) and the Netherlands Organisation for Applied Scientific Research (TNO), 2018, https://doi.org/10.3997/2214-4609.201802288, https://www.earthdoc.org/content/papers/10.3997/2214-4609.201802288.

[40] J. Owen, *Bayesian Uncertainty Analysis and Decision Support for Complex Models of Physical Systems with Application to Production Optimisation of Subsurface Energy Resources*, 2022, http://etheses.dur.ac.uk/14751/.

[41] J. Owen, I. Vernon, and J. Carter, *Bayesian Emulation of Complex Computer Models with Structured Partial Discontinuities*, vol. 435 of Springer Proceedings in Mathematics & Statistics, Springer, 2023, pp. 1–13, https://doi.org/10.1007/978-3-031-42413-7.

[42] J. Owen, I. Vernon, and R. Hammersley, *A Bayesian Statistical Approach to Decision Support for TNO OLYMPUS Well Control Optimisation under Uncertainty*, in ECMOR XVII – 17th European Conference on the Mathematics of Oil Recovery, vol. 2020, European Association of Geoscientists & Engineers (EAGE), 2020, https://doi.org/10.3997/2214-4609.202035109.

[43] M. Plumlee and R. Tuo, *Building accurate emulators for stochastic simulations via quantile kriging*, Technometrics, 56 (2014), pp. 466–473, https://doi.org/10.1080/00401706.2013.860919, https://www.tandfonline.com/doi/full/10.1080/00401706.2013.860919.

[44] F. Pukelsheim, *The Three Sigma Rule*, The American Statistician, 48 (1994), pp. 88–91, https://doi.org/10.2307/2684253.

[45] M. Raissi, P. G. Perdikaris, and G. E. Karniadakis, *Physics-informed neural networks: A deep*

learning framework for solving forward and inverse problems involving nonlinear partial differential equations, Journal of Computational Physics, 378 (2019), pp. 686–707, https://doi.org/10.1016/j.jcp.2018.10.045, https://www.sciencedirect.com/science/article/pii/S0021999118307125.

[46] V. Rannou, F. Brouaye, M. Hélier, and W. Tabbara, *Kriging the quantile: application to a simple transmission line model*, Inverse Problems, 18 (2002), pp. 37–48, https://doi.org/10.1088/0266-5611/18/1/303, https://dx.doi.org/10.1088/0266-5611/18/1/303.

[47] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, The MIT Press, 2006, https://gaussianprocess.org/gpml/chapters/RW.pdf.

[48] J. Rougier, M. Goldstein, and L. House, *Second-Order Exchangeability Analysis for Multimodel Ensembles*, Journal of the American Statistical Association, 108 (2013), pp. 852–863, https://doi.org/10.1080/01621459.2013.802963.

[49] J. Rougier, S. Guillas, A. Maute, and A. D. Richmond, *Expert Knowledge and Multivariate Emulation: The Thermosphere–Ionosphere Electrodynamics General Circulation Model (TIE-GCM)*, Technometrics, 51 (2009), pp. 414–424, https://doi.org/10.1198/TECH.2009.07123.

[50] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn, *Design and Analysis of Computer Experiments*, Statistical Science, 4 (1989), https://doi.org/doi:10.1214/ss/1177012413.

[51] T. J. Santner, B. J. Williams, and W. I. Notz, *The Design and Analysis of Computer Experiments*, Springer Series in Statistics, Springer-Verlag, 1st ed., 2003, https://doi.org/10.1007/978-1-4757-3799-8.

[52] R. Schulze-Riegert, A. Anton, J. Baffoe, D. Geissenhoener, K. J. Ng, M. Nwakile, S. Skripkin, C. Meulengracht, and M. Whymark, *Standardized workflow design for field development plan optimization under uncertainty*, European Association of Geoscientists and Engineers (EAGE) and the Netherlands Organisation for Applied Scientific Research (TNO), 2018, https://doi.org/10.3997/2214-4609.201802290, https://www.earthdoc.org/content/papers/10.3997/2214-4609.201802290.

[53] C. Tebaldi and R. Knutti, *The use of the multi-model ensemble in probabilistic climate projections*, Philosphical Transactions of the Royal Society A, 364 (2007), pp. 2053–2075, https://doi.org/10.1098/rsta.2007.2076, https://royalsocietypublishing.org/doi/pdf/10.1098/rsta.2007.2076.

[54] TNO, *OLYMPUS Oil Reservoir Model Input Decks*, 2017, https://www.isapp2.com/downloads/olympus-reservoir-model.pdf.

[55] TNO, *Integrated Systems Approach for Petroleum Production (ISAPP) Optimisation Challenges*, 2018, http://www.isapp2.com/home.html.

[56] I. Vernon, M. Goldstein, and R. Bower, *Galaxy Formation: Bayesian History Matching for the Observable Universe*, Statistical Science, 29 (2014), pp. 81–90.

[57] I. Vernon, M. Goldstein, and R. G. Bower, *Galaxy Formation: a Bayesian Uncertainty Analysis*, Bayesian Analysis, 5 (2010), pp. 619–670.

[58] I. Vernon, S. E. Jackson, and J. A. Cumming, *Known Boundary Emulation of Complex Computer Models*, SIAM/ASA Journal on Uncertainty Quantification, 7 (2019), pp. 838–876, https://doi.org/10.1137/18M1164457.

[59] I. Vernon, J. Liu, M. Goldstein, J. Rowe, J. Topping, and K. Lindsey, *Bayesian uncertainty analysis for complex systems biology models: emulation, global parameter searches and evaluation of gene functions*, BMC systems biology, 12 (2018), https://doi.org/10.1186/s12918-017-0484-3.

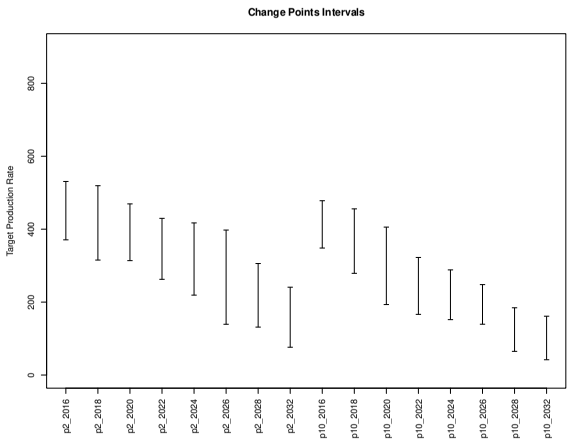
[60] I. Vernon, J. Owen, J. Aylett-Bullock, C. Cuesta-Lazaro, J. Frawley, A. Quera-Bofarull, A. Sedgewick, D. Shi, H. Truong, M. Turner, J. Walker, T. Caulfield, K. Fong, and F. Krauss, *Bayesian Emulation and History Matching of JUNE*, Philosphical Transactions of the Royal Society A, 380 (2022), https://doi.org/10.1098/rsta.2022.0039.

[61] I. Vernon, J. Owen, and J. Carter, *Bayesian Emulation for Computer Models with Multiple Partial Discontinuities*, Bayesian Analysis, (2024), https://doi.org/10.1214/24-BA1456, https://doi.org/10.1214/24-BA1456, https://arxiv.org/abs/2210.10468v1.

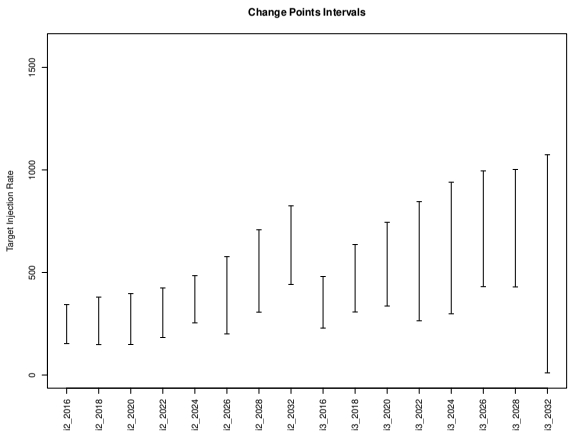
[62] D. F. Vysochanskij and Y. I. Petunin, *Justification of the 3-σ Rule for Unimodal Distribution*, Theory of Probability and Mathematical Statistics, 21 (1980), pp. 25–36.

[63] D. Williamson, M. Goldstein, L. Allison, A. Blaker, P. Challenor, L. Jackson, and K. Yamazaki, *History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble*, Climate Dynamics, 41 (2013), pp. 1703–1729,

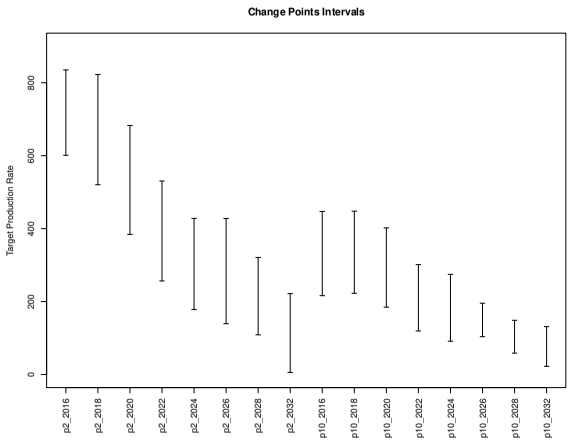
https://doi.org/10.1007/s00382-013-1896-4.
[64] D. Williamson, M. Goldstein, and A. Blaker, *Fast linked analyses for scenario-based hierarchies*, Journal of the Royal Statistical Society: Series C (Applied Statistics), 61 (2012), pp. 665–691, https://doi.org/10.1111/j.1467-9876.2012.01042.x.



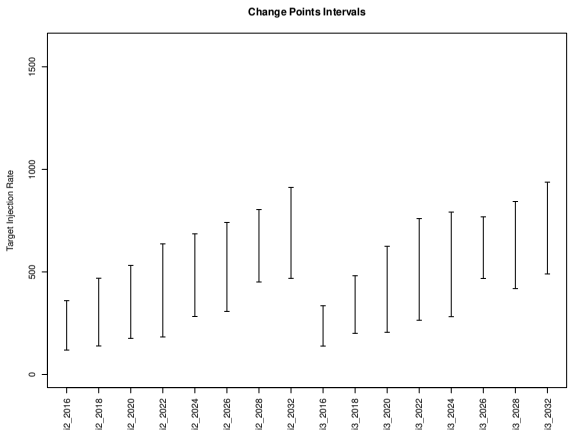
(a) OLYMPUS 25 WOPT



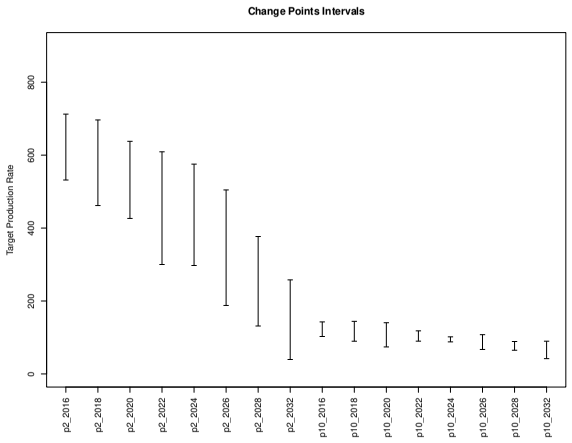
(b) OLYMPUS 25 WWIT



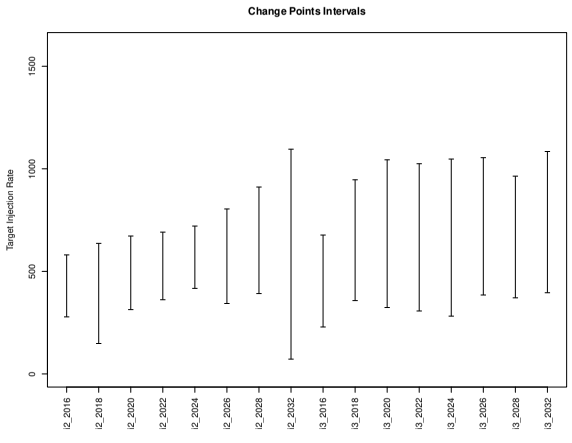
(c) OLYMPUS 33 WOPT



(d) OLYMPUS 33 WWIT



(e) OLYMPUS 45 WOPT



(f) OLYMPUS 45 WWIT

Figure 15: OLYMPUS wave 1 change point upper bound and extrapolation cut-off intervals for WOPT and WWIT within each control interval with respect to their corresponding decision parameter for each of the three sub-sampled OLYMPUS models.