Proceedings of the ASME 2020 39th International Conference on Ocean, Offshore and Arctic Engineering
OMAE 2020
June 28-July 3, 2020, Fort Lauderdale, FL, USA

OMAE2020-19119

OFFSHORE DRILLING: EXTENDING THE WEATHER WINDOW FOR OPERATIONS BY OPTIMAL USE OF SIMULATIONS AND PROBABILISTIC MACHINE LEARNING

Simen Eldevik*
Group Technology and Research
DNV GL
Oslo, Norway
Email: simen.eldevik@dnvgl.com

Stian Sætre, Erling Katla, Andreas B. Aardal
Riser Technology, Oil & Gas
DNV GL
Oslo, Norway

ABSTRACT

Operators of offshore floating drilling units have limited time to decide on whether a drilling operation can continue as planned or if it needs to be postponed or aborted due to oncoming bad weather. With day-rates of several hundred thousand USD, small delays in the original schedule might amass to considerable costs. On the other hand, pushing the limits of the load capacity of the riser-stack and wellhead may compromise the integrity of the well itself, and such a failure is not an option.

Advanced simulation techniques may reduce uncertainty about how different weather scenarios influence the system's integrity, and thus increase the acceptable weather window considerably. However, real-time simulations are often not feasible and the stochastic behavior of wave-loads make it difficult to simulate all relevant weather scenarios prior to the operation.

This paper outlines and demonstrates an approach which utilizes probabilistic machine learning techniques to effectively reduce uncertainty. More specifically we use Gaussian process regression to enable fast approximation of the relevant structural response from complex simulations. The probabilistic nature of the method adds the benefit of an estimated uncertainty in the prediction which can be utilized to optimize how the initial set of relevant simulation scenarios should be selected, and to predict real-time estimates of the utilization and its uncertainty when combined with current weather forecasts.

This enables operators to have an up-to-date forecast of the system's utilization, as well as sufficient time to trigger additional scenario-specific simulation(s) to reduce the uncertainty of the current situation. As a result, it reduces unnecessary conservatism and gives clear decision support for critical situations.

Keywords: Offshore drilling, Weather window, Simulations, Design-of-experiments, Probabilistic machine learning, Gaussian process regression, Uncertainty.

INTRODUCTION

Offshore drilling is a complex and expensive operation that may incur additional costs of several hundred thousand USD per day due to weather and interrupted or postponed operations. Prior to and during any offshore operation, the operator assesses the potential for interruptions based on current weather and ocean forecasts, previous experience, and simulated resilience of the rig, the riser stack-up, and the subsea wellhead.

As the operator has limited time between a weather forecast is available and when a decision needs to be taken with respect to continued or aborted operation, the efficiency of acquiring accurate information for a well-informed decision is critical to avoid unnecessary interruption or delays. It is often necessary to assess and establish operational limits for specific operations where the integrity of the riser or well equipment is the limiting factor. This is a complex process where specific rig, riser, and well information is used to establish several dynamic models that need to be assessed in relation to one another. In this

*Contact author: simen.eldevik@dnvgl.com

paper we investigate a system where one model has been used to explore the rig's response to different weather scenarios and sea-states, while another is needed to assess the structural integrity and utilization of the riser and the riser-well connection given the dynamic response of the rig. These models are complex and time-consuming, and thus, it is a computational challenge to provide real-time decision support.

A common approach is to run a lot of these simulations prior to the operation, and to establish conservative thresholds for operational windows based on worst-case combinations of parameters [1].

There are several drawbacks with this approach:

1. **A priori selection of simulation scenarios** might result in a significant portion of the simulation effort being spent in regions of the response space which are either non-consequential (i.e. we have a large margin to the critical utilization level) or heavily over-utilized (i.e. far beyond the critical utilization level).
2. **Dynamic and stochastic nature of the current situation** makes it difficult to assess the forecasted sea-state change (including uncertainties) with all the relevant simulation results (and whether all relevant simulations have been run).
3. This often result in a **qualitative judgement** of the worst sea-state compared to the assumed most relevant simulation result(s) – a cumbersome decision process.

When high-fidelity and real-time results for specific scenarios are not available, the safety philosophy is to be conservative in light of the uncertainty. This paper presents an approach which mitigate the shortcomings listed above and reduce uncertainty relevant to the decision context, and thus the need to be overly conservative.

The current work is based on results presented by Eldevik and Sætre in [1] which demonstrates how probabilistic machine learning can be used to optimize simulation efforts prior to operations. The results presented in this paper is a direct continuation of that work, and uses the established relation between the environmental loads, the response of the rig and the response and capacity of the riser-well system to enhance the decision support of the operator in real-time based on current weather forecasts.

SYSTEM DESCRIPTION

For the system under consideration, we want to continuously assess the structural integrity of the riser stack and the wellhead for relevant weather scenarios. This require advanced and time-consuming simulations of the dynamic movement of the drill rig for all potential sea-states it can experience during operation (i.e. wave loads and current), its effect on the structural response of the riser system and wellhead and the system's capacity as discussed in [1]. The utilization of this capacity u, is treated as an unknown function f of a set of relevant parameters $\boldsymbol{x}$

$$u = f(\boldsymbol{x}) \tag{1}$$

The parameters that dominate the utilization are the sea-state, described by significant wave height, H_s, wave period, T_p, and the current velocity, v_c, the dynamic position and off-set from the center, d, and the internal pressure of the riser, p_{int}.

The function f is a mapping of the five-dimensional input space to the one-dimensional output space, $f : \mathbb{R}^5 \to \mathbb{R}$, and the challenge is to explore this multi-dimensional input space sufficiently to have the necessary confidence in the estimated utilization for all relevant operation- and weather-scenarios.

This paper briefly describes how this exploration has been done in [1] to optimize the simulation effort for input scenarios close to the critical , i.e. continued or aborted drilling operation. Then it continues to expand on this work by suggesting an approach where additional simulations can be triggered and run live to reduce unnecessary aborted operations while maintaining the necessary confidence that the rig and riser-well system can withstand the forecasted weather scenarios.

SURROGATE MODELS AND ADAPTIVE EXPLORATION

Generally, if f is a time-consuming and computationally expensive simulation or costly experiment, it is not possible to practically evaluate $f(\boldsymbol{x})$ for all relevant $\boldsymbol{x}$ within the time-frame of the relevant decision context. Thus, we create a fast-running approximate model $\hat{f} \approx f$ to the real response based on a number of evaluations $\{f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_N)\}$, which are able to predict the response also away from the observed (training) data. This is what most machine learning (ML) algorithms produce - and is often referred to as surrogate models.

However, for safety-critical decisions, it is important to be able to quantify the uncertainty and sensitivity related to predictions from such a surrogate model. Gaussian Process (GP) regression is a probabilistic machine learning model that have two important traits that support this. A GP can:

interpolate (i.e. make sure that all predictions $\hat{f}(\boldsymbol{x}_i) = f(\boldsymbol{x}_i)$ match the observations at the training points $\{\boldsymbol{x}_i\}_{i=1}^N$), and

provide an uncertainty estimate on $\hat{f}(\boldsymbol{x})$ for all predictions away from the training data $\boldsymbol{x} \notin \{\boldsymbol{x}_i\}_{i=1}^N$.

O'Hagan [2] refer to surrogate models having these traits as emulators.

To establish an emulator that can be used for decision support, we thus need to evaluate or observe the expensive function f a finite number of times. Engineers have a tendency to decide on a factorial approach [3] which explore the input space in a grid-like manner. This approach is known to be inferior, but is nevertheless often used because of its simplicity [4]. Eldevik and

Sætre [1] also show how such an approach is ignorant of the critical limit of the response, and, thus, spend a significant amount of the simulation efforts in areas of the state space that have an insignificant effect on the decision problem (i.e. where a lot of the simulation effort has been spent where the system is well within, or far outside, the system's capacity limit). To mitigate this [1] utilize an adaptive approach to focus the simulation effort close to the critical decision limit.

PROBABILISTIC ML - GAUSSIAN PROCESS

As we will use the results and model from [1] the theoretical description of the GP regression model is paraphrased in this section. Rassmussen and Williams [5] define the mean and covariance function of a real process $f(\boldsymbol{x})$ as

$$m(\boldsymbol{x}) = \mathbb{E}\left[f(\boldsymbol{x})\right] \tag{2a}$$

$$k(\boldsymbol{x},\boldsymbol{x}') = \mathbb{E}\left[\left(f(\boldsymbol{x}) - m(\boldsymbol{x})\right)\left(f(\boldsymbol{x}') - m(\boldsymbol{x}')\right)\right] \tag{2b}$$

and write the Gaussian Process as

$$f \sim \mathcal{GP}(m(\boldsymbol{x}), k(\boldsymbol{x},\boldsymbol{x}')). \tag{3}$$

The GP is completely specified by this mean and covariance functions, and describes a distribution over functions. For any finite collection $[\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n]$ of points $\boldsymbol{x}_i \in \mathbb{R}^D$, the distribution of $[f(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)]$ is multivariate Gaussian with mean $[m(\boldsymbol{x}_1),\ldots,m(\boldsymbol{x}_m)]$ and covariance $\mathbf{K}_{i,j} = k(\boldsymbol{x}_i,\boldsymbol{x}_j)$, and the marginal distribution of a subsets of the random variables will also be Gaussian.

Thus, for a data set $\mathcal{D}$ of N noisy observations of the real process, i.e. $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^N$ where $y_i = f(\boldsymbol{x}_i) + \varepsilon_i$ and we assume the noise term ε_i is i.i.d. normally distributed with zero-mean and common variance c^2, we can infer the distribution of $f(\boldsymbol{x}^*)|\mathcal{D}$ at unobserved points $\boldsymbol{x}^*$ in the input space. The predictive posterior distribution at n new inputs $\boldsymbol{f}^*|\mathcal{D} = [f(\boldsymbol{x}_1^*),\ldots,\boldsymbol{x}_n^*]|\{y_i = f(\boldsymbol{x}_i)+\varepsilon_i\}_{i=1}^N$ is given by

$$\boldsymbol{f}^*|\mathcal{D} \sim N\left(\mu_{\boldsymbol{f}^*|\mathcal{D}}, \Sigma_{\boldsymbol{f}^*|\mathcal{D}}\right) \tag{4}$$

with mean and covariance

$$\mu_{\boldsymbol{f}^*|\mathcal{D}} = \boldsymbol{m}_* + \boldsymbol{K}_*\left(\boldsymbol{K}_* + c^2\boldsymbol{I}\right)^{-1}(\boldsymbol{y} - \boldsymbol{m}) \tag{5a}$$

$$\Sigma_{\boldsymbol{f}^*|\mathcal{D}} = \boldsymbol{K}_{**} - \boldsymbol{K}_*\left(\boldsymbol{K} - c^2\boldsymbol{I}\right)^{-1}\boldsymbol{K}_*^T \tag{5b}$$

where $\boldsymbol{m}_*$, $\boldsymbol{m}$ and $\boldsymbol{y}$ are vectors with elements $m(\boldsymbol{x}_i^*)$, $m(\boldsymbol{x}_i)$ and y_i respectively. $\boldsymbol{I}$ is the $N \times N$ identity matrix, and $\boldsymbol{K}_*$ and $\boldsymbol{K}_{**}$ have elements $(\boldsymbol{K}_*)_{i,j} = k(\boldsymbol{x}_i^*, \boldsymbol{x}_j)$ and $(\boldsymbol{K}_{**})_{i,j} = k(\boldsymbol{x}_i^*, \boldsymbol{x}_j^*)$ respectively.

The GP will approximate the utilization function, and because the critical decision limit is $u = 1-0$ it is reasonable to select a prior mean function $m(\boldsymbol{x}) = 1.0$, so that the model will revert to suggest unacceptable utilization when it does no have sufficient evidence to suggest otherwise. The covariance function k governs how much the function can vary between two points that are close to each other. The Mátern 3/2 kernel, which guarantees that the function is once differentiable, has been used in [1], and is defined by

$$k(x,x') = \sigma^2\left(1+\sqrt{3}r\right)\exp\left(-\sqrt{3}r\right), \tag{6}$$

where

$$r = \sqrt{\sum_{i=1}^{\mathcal{D}} \frac{(x_i - x_i')^2}{l_i}}. \tag{7}$$

σ is the kernel variance and l_i is the length scale in dimension i.

The GP model is trained by estimating the hyperparameters θ, which comprises: the kernel variance σ, the length scale parameters l_i and the white noise variance c, based on the training data, typically by maximum likelihood or Bayesian methods [5]. We want to find the set of hyperparameters θ_{max} which maximizes the probability of observing the training data in the GP model.

In this work, the *Pyro: Deep Universal Probabilistic Programming* language [6] has been used to train the GP regression model and make predictions according to Eqn. (5a) and Eqn. (5b). A detailed tutorial on how to train a GP within this framework can be found on http://pyro.ai/examples/gp.html

OPTIMIZE ACCORDING TO DECISION CONTEXT

As shown in [1], the Bayesian nature of the GP model is very suitable for exploration-exploitation strategies based on uncertainty reduction. The GP model established in [1] was established by exploring the structural response and utilization of critical components within the riser stack up according to the method described by Ranjan et al. [7], focusing on system responses that correspond to the critical utilization limit $u = 1.0$.

For the operators, it is important to know where the boundary between acceptable and unacceptable operating conditions are. As Eldevik and Sætre state in [1] "We do not want the GP model to support a decision to abort an operation when in fact it could have continued safely, but more importantly, we cannot let the GP model support a decision to continue operation when in fact it will lead to a structural failure of the riser-well system."

The method applied focused the simulation efforts in the vicinity of the critical response utilization $u = 1.0$. Less effort was spent on scenarios where the utilization was far away from this decision boundary - e.g. in very calm sea-states or in very bad weather scenarios where the decision is often trivial. See Eldevik and Sætre [1] for more details on the adaptive approach for optimal reduction of uncertainty used as basis in this work.

Figure 1a shows the prediction from a GP model trained on 500 adaptively selected simulations [1], and the rest of this section is a summarization of these results. 5000 simulations of random weather scenarios have been used to test the efficacy of the GP model across the entire input-space. The predictions from the GP model is plotted as blue dots on the y-axis versus the actual simulation result on the x-axis. Dots that are on the dashed unity line signify perfect predictions, while the spread around the dashed unity line illustrate the uncertainty in the GP model (i.e. either conservative over prediction or non-conservative under prediction by the GP model).

As the decision context is related to the binary outcome where the prediction is either above or below the critical limit this can be thought of as a classification problem where a *Positive* prediction is where the operator decide to commence or continue the drilling operation, while a *Negative* prediction is where the operator decide to postpone or abort the drilling operation.

For low utilization (See Fig. 1a below the critical limit in the lower left corner) the weather scenarios are within the operational window of the drill rig. These are true positive predictions and far below the limit the outcome of the decision is OK irrespective of the uncertainty in the GP model. Similarly, above the critical limit (in the upper right corner), the utilization is more than the system can withstand and the decision to discontinue the operation is the correct one. These are true negative predictions are far above the limit uncertainty in the GP prediction do not affect the outcome of the decision.

The critical outcomes are where the under- or over-prediction may lead to a critical False Positive or costly False Negative decision, and this depends on the magnitude of the predicted utilization.

For those scenarios where the GP prediction is above the decision criteria, and thus the operation is stopped, while the true simulated condition was OK leads to a False Negative decision. These outcomes will be in the upper left corner, shaded yellow, and signify an unnecessary disconnection and abortion of the operation – incurring additional and unnecessary costs. The most critical erroneous decisions are those where the prediction suggest that it is OK to continue operation, but where this is not the case. This leads to a False Positive decision and is illustrated by predictions that fall in the lower right corner shaded dark red.

With respect to the decision context, the objective of the work in [1] was to ensure that the rate of False Positive decisions from the GP model close to ~ 0, while minimizing the rate of the less critical, but costly, False Negative decisions. To achieve this the uncertainty in the GP model need to be small close to the decision boundary (small spread of blue dots), while it can accept a larger spread further from the decision boundary.

Note that all the GP predictions have been moved ”upward” in the plot corresponding to a 95% confidence level (i.e. 1.645 standard deviation). This has been done to minimize the possibility of the GP model to make a critical under-prediction (red lower right), while the focused simulation effort close to the critical limit ensure that this shift does not lead to a lot of unnecessary aborted operations because of overly conservative predictions (orange upper left).

Figure 1b shows the corresponding confusion matrix to the decision outcomes discussed above. The numbers shows the fraction of GP predictions which lead to correct and erroneous decisions. See [1] for details on how the optimization of simulation efforts has been done when training the GP model to ensure safe operations with $\sim 0\%$ critical False Positive decisions and only 0.9% conservative False Negative decisions.

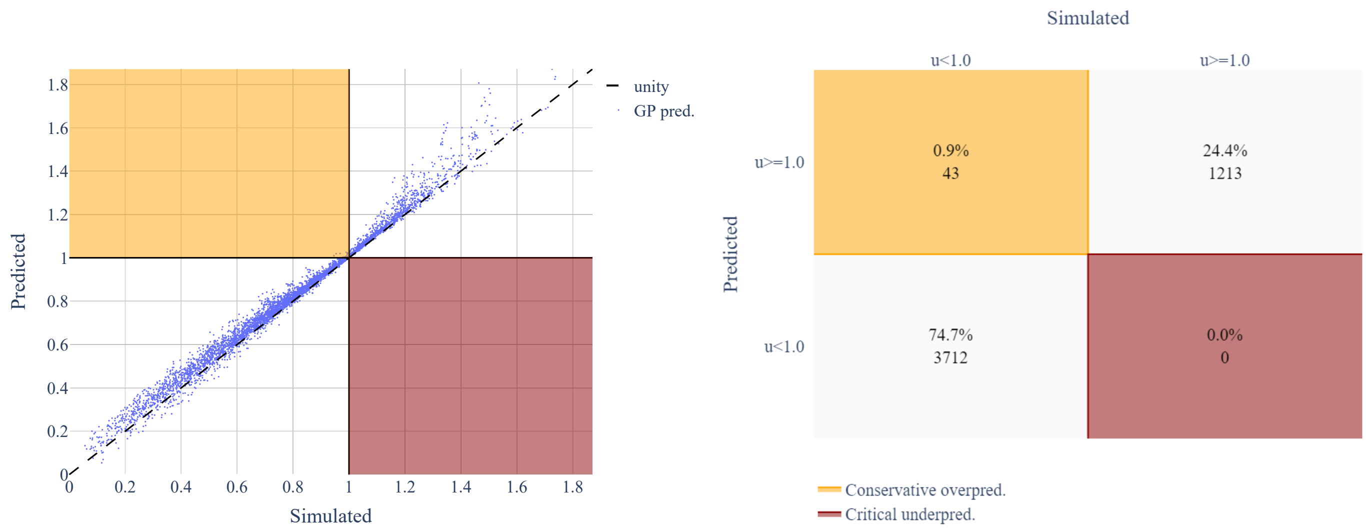
Note that it is imperative that machine-learned models used for safety-critical systems incorporate a measure of risk, i.e. can quantify the risk of making erroneous predictions [8]. The ability of probabilistic methods like GP models to quantify uncertainty is key to enable such a risk measure. In this work, the risk measure is the uncertainty related to making the wrong decision compared to the consequence of making that erroneous decision, i.e. the False Positive or False Negative decisions.

As noted in [1] the 5000 samples shows a reasonable uncertainty level for all the sampled weather scenarios, but this do not guarantee that the model is sufficiently good for all weather conditions that might be experienced in the future. In the next section we expand on how these results can be used by the operator in the dynamic decision context they continuously need to assess. As a GP model is fast and computationally inexpensive, it can act as a real-time support tool when combined with online forecast services and on-demand simulation capabilities. And, it can be used to guide the selection of specific simulation scenarios to reduce the uncertainty in forecasted utilization that is not sufficiently low based on the current knowledge base.

DYNAMIC DECISION CONTEXT

Weather and sea states are stochastic in nature, and it is difficult to have long-term forecasts with the necessary accuracy for good decision support prior to any drilling operation of some length. However, the forecasts are usually sufficiently accurate for the next 24 hour period, and this time horizon is long enough to run a few simulations or prepare a disconnect of the riser from the well. Figure 2 shows a typical marinogram from the Norwegian online weather forecast service www.yr.no downloaded December 9th 2019.

The forecasted sea state parameters for the next four days and typical operational pressure of 31 MPa and a positioning limitation of 14 m offset is inputted to the GP model to produce a utilization forecast as shown in Fig. 3. The GP model outputs an hourly mean prediction (dark blue line denoted ”mu”) and



(a) Prediction vs. simulation

(b) Confusion matrix

FIGURE 1: GP PREDICTION VS. SIMULATION RESULTS OF UTILIZATION [1].

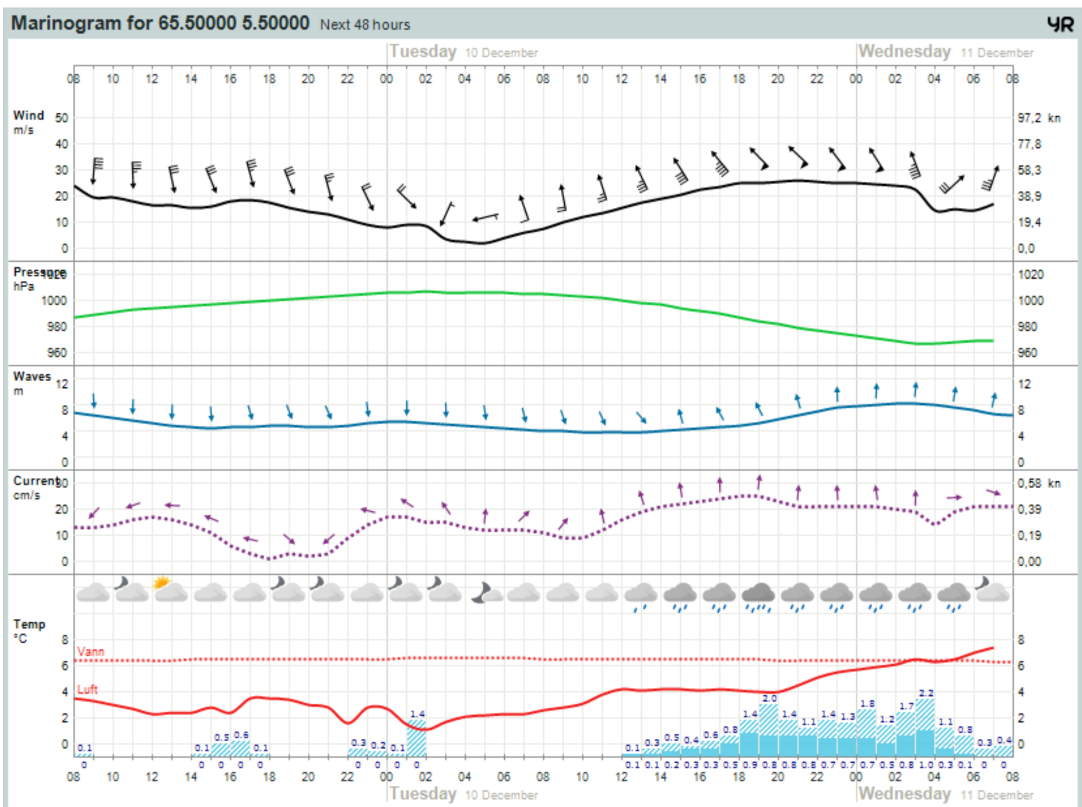


FIGURE 2: MARINOGRAM FROM WEATHER SERVICE WWW.YR.NO.

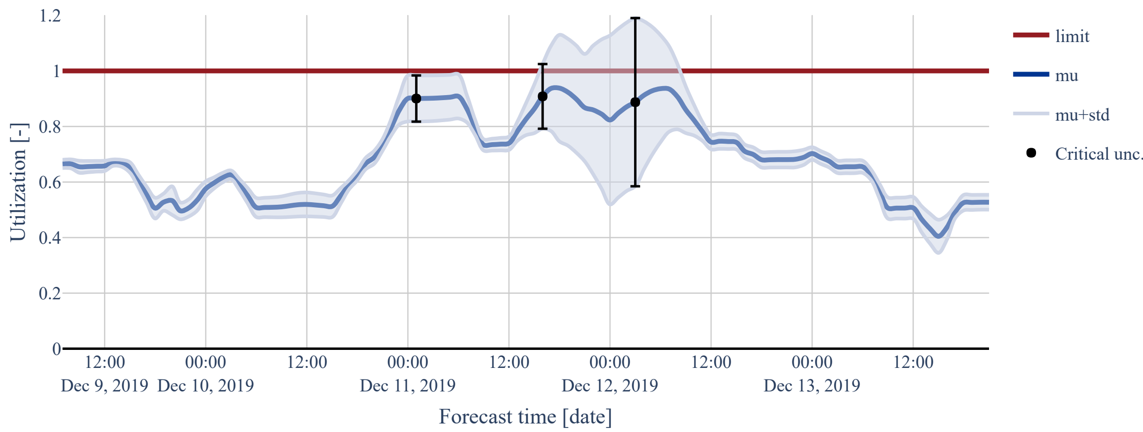


FIGURE 3: GP PREDICTION OF UTILIZATION FOR THE GIVEN WEATHER FORECAST.

its associated standard deviation (light blue shaded area denoted "mu+std") of the utilization. The red solid line denoted "limit" is the critical utilization threshold of $u = 1.0$. This forecast of utilization corresponding to the relevant forecasted weather-states makes it possible to have a much more dynamic view of the operational limitations than current approaches which are, to a large extent, based on worst case combinations of input data.

The figure also mark three points in the utilization forecast where the uncertainty in the prediction affects the decision significantly, and where reduced uncertainty would affect this decision the most (black dots). As this example shows, the mean utilization prediction is just below the critical threshold for the entire forecast. However, at three times the uncertainty in the prediction is so large that a significant portion of the distribution is above the threshold. This increase in uncertainty is most likely due to a lack of simulations results for the relevant combinations of weather data, rig offset and internal riser pressure. As stated initially, it is not possible to run simulations for all possible input combinations (i.e. evaluate $f(\boldsymbol{x})\forall\boldsymbol{x}$), and thus some weather scenarios will not have been evaluated prior to commencing operation. No matter which exploration strategy is used to establish the initial data set (e.g. factorial approaches, random approaches, or adaptive approaches) there will be parts of the input space that has not been explored and where the surrogate model must extrapolate. These weather scenarios will, in a Bayesian framework, have a larger associated uncertainty.

EFFECT OF REDUCING UNCERTAINTY

The benefit of Bayesian methods, compared to ML methods that do not provide an uncertainty estimate, is that it is able to identify scenarios with higher uncertainty. Based on this, it is possible, during operations, to update the data set with simulations which specifically probe the forecasted weather scenarios with the high uncertainty close to the critical limit. This is a significant advantage compared to the current practice where all simulations are done prior to commencing operation.

If we look at the three points in Fig. 3 marked as having a critical uncertainty, these points have been selected based on a criteria where the upper estimate of the utilization including uncertainty ("mu+std") is close to or above the critical limit ($> u_{limit} - \varepsilon_{ub}$) while there still is a significant portion of the distribution that predicts an acceptable utilization ($< u_{limit} - \varepsilon_{lb}$). This can be expressed formally by:

$$\mu_{\boldsymbol{f}^*|\mathcal{D}} + \Sigma_{\boldsymbol{f}^*|\mathcal{D}} > u_{limit} - \varepsilon_{ub} \quad \text{and} \tag{8a}$$

$$\mu_{\boldsymbol{f}^*|\mathcal{D}} - \Sigma_{\boldsymbol{f}^*|\mathcal{D}} < u_{limit} - \varepsilon_{lb} \tag{8b}$$

Here ε_{ub} controls how close to the critical limit an upper bound (ub) of the predicted utilization should be, and ε_{lb} controls how far below the critical limit a lower bound (ub) of the predicted utilization should be, to justify spending simulation effort at this weather scenario.

If we can reduce the uncertainty for these weather scenarios in time for them to occur, we might be able to continue operating even though the initial prediction supports a decision to disconnect. The first of the three selected weather scenarios are 36 hours from the time of the forecast. This is enough to run a simulation for these specific weather scenarios, and thus potentially reduce the uncertainty of the utilization prediction, increasing the confidence in the decision to either disconnect and abort or continue the operation.

Utilizing the Bayesian nature of the trained GP model, we can test how much the uncertainty will be affected by getting one or more additional simulation results. If we assume that the three identified scenarios with the highest associated uncertainty would result in the mean value predicted by the GP model, we can update the model with these observations, and get a quick estimate of how the uncertainty of the entire forecast would be affected. Figure 4 shows the effect of adding one, two, or three *hypothetical* simulation results.

Note that these *hypothetical* simulation results have not yet actually been run, and the uncertainty reduction is only illustrating how the overall uncertainty may be affected by added information.

In this example, it is clear that without simulating additional weather scenarios, the uncertainty in the predicted utilization for the forecasted weather suggests that the rig need to disconnect from the well. However, it also suggests that if three specific weather scenarios can be simulated within the next 24 hours, this might produce the evidence needed to be able to continue the drilling operations with necessary confidence.

CONCLUDING REMARKS

In this work we have showed the efficacy of probabilistic machine learning, and more specifically Gaussian Process regression, with respect to guiding optimal simulation efforts and real-time decision support for offshore drilling operations. As this is a safety-critical operation, an uncertainty measure is essential when applying data-driven methods in real-time decision support.

- Gaussian Process regression is well suited as an emulator for complex simulations of critical response.
- The Bayesian nature of GP models and its uncertainty estimate is well suited to implement in the objective function of information-based sequential data exploration. This enables focused simulation efforts towards consequential scenarios rather than spending a lot of simulation effort in non-consequential scenarios.
- Due to the reduced uncertainty close to the critical decision criteria, an upper confidence bound of the GP model prediction enables the engineer to reduce the probability of critical under-prediction while at the same time limit its effect on overly conservative predictions.
- Combining a trained GP model with real-time operational data and forecasted weather scenarios gives a powerful and easy-to-use decision support for the operator.
- Clear communication of where the uncertainty of the model prediction is high can be used to trigger simulations that specifically targets the uncertain scenarios. This might provide the evidence needed to operate through weather that previously would have called for an aborted operation.

It is clear that added information will result in higher confidence in critical decisions. Making the relevant information available to the decision maker in clear terms increases the potential for optimal decision making, and reduce the number of unnecessary disconnections.

Note, however, that these results only illustrate how an optimal decision support may be achieved, and that the confidence of the decision also needs to take into account the accuracy of the simulation model(s) and weather forecasts in addition to the prediction uncertainty.

ACKNOWLEDGMENT

Thanks to Xu Han for considerable simulation efforts that have made this project possible and DNV GL for supporting the work.

REFERENCES

[1] Eldevik, S., 2020. “Offshore workover operations: Reducing uncertainty of critical weather scenarios by optimal use of simulations and probabilistic machine learning”. In Proceedings of the 30th European Safety and Reliability Conference and the 15th Probabilistic Safety Assessment and Management Conference, P. Baraldi, F. D. Maio, and E. Zio, eds., European Safety and Reliability Association, Research Publishing. Paper number 4901 - *preprint*.

[2] O’Hagan, A., 2006. “Bayesian analysis of computer code outputs: A tutorial”. *Reliability Engineering & System Safety,* ***91***(10-11), pp. 1290–1300.

[3] Fisher, R. A., 1937. *The Design of Experiments*, 2nd ed. Oliver and Boyd.

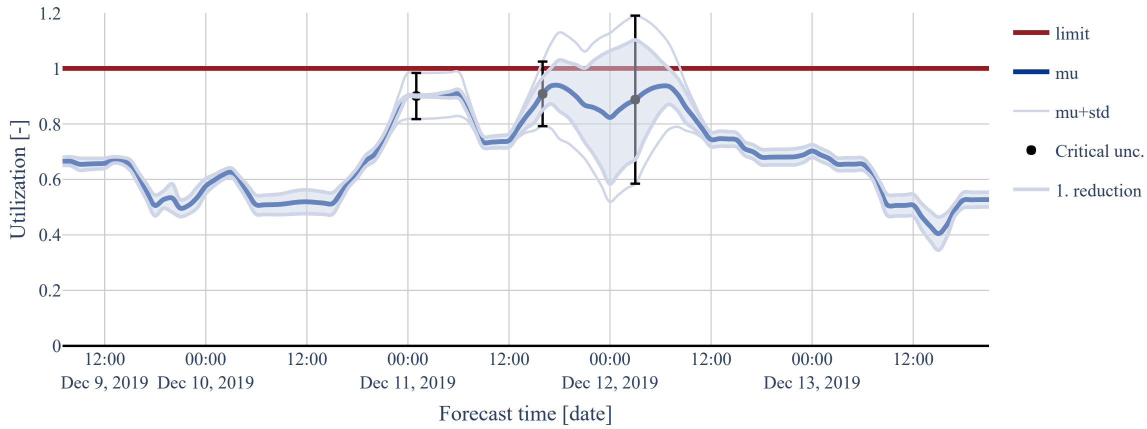
[4] Saltelli, A., and Annoni, P., 2010. “How to avoid a perfunctory sensitivity analysis”. *Environmental Modelling & Software,* ***25***(12), December, pp. 1508–1517.

[5] Rasmussen, C. E., and Williams, C. K. I., 2006. *Gaussian Processes for Machine Learning*. The MIT Press.

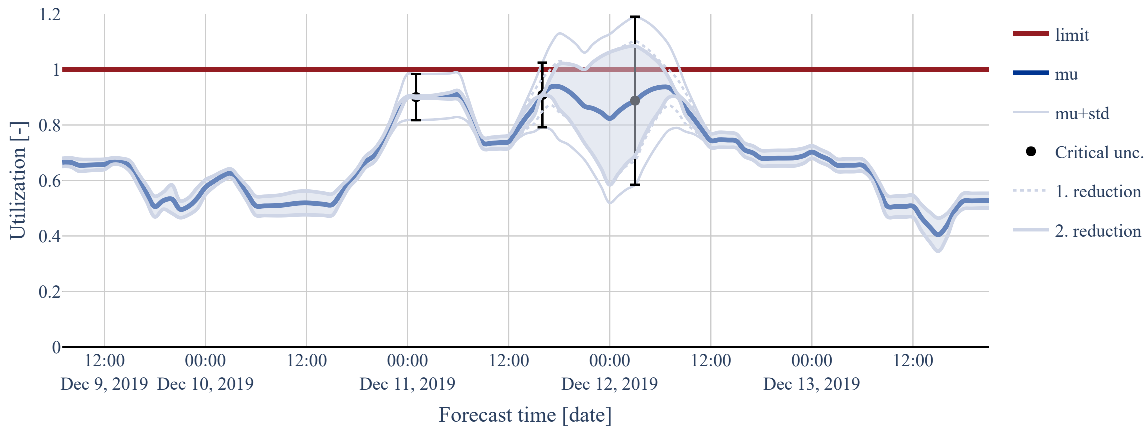
[6] Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., and Goodman, N. D., 2018. “Pyro: Deep Universal Probabilistic Programming”. *arXiv preprint arXiv:1810.09538*.

[7] Ranjan, P., Bingham, D., and Michailidis, G., 2008. “Sequential experiment design for contour estimation from complex computer codes”. *Technometrics,* ***50***(4), pp. 527–541.

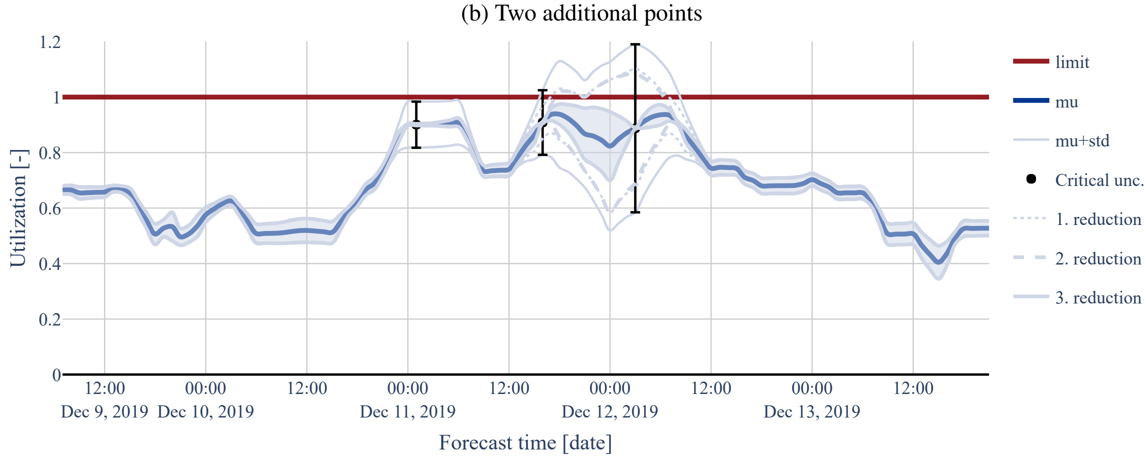
[8] Eldevik, S., 2018. Ai + safety. Tech. rep., Group Technology & Research, DNV GL, Høvik, Norway, August. Position paper available at https://ai-and-safety.dnvgl.com/.



(a) One additional point



(b) Two additional points



(c) Three additional points

FIGURE 4: EFFECT ON UNCERTAINTY IN PREDICTION IF THREE MORE SIMULATIONS WHERE RUN FOR SCENARIOS WITH HIGH UNCERTAINTY