Bayesian Reliability Assessment of Legacy Safety-Critical Systems Upgraded with Fault-Tolerant Off-the-Shelf Software

Peter Popov
Centre for Software Reliability, City University,
Northampton Square, London, UK
e-mail: ptp@csr.city.ac.uk
phone: +44 207 040 8963
fax: +44 207 040 8585

Abstract. This paper presents a new way of applying Bayesian assessment to systems, which consist of many components. Full Bayesian inference with such systems is problematic, because it is computationally hard and, far more seriously, one needs to specify a multivariate prior distribution with many counterintuitive dependencies between the probabilities of component failures. The approach taken here is one of *decomposition*. The system is decomposed into *partial views* of the systems or part thereof with different degrees of detail and then a mechanism of propagating the knowledge obtained with the more refined views back to the coarser views is applied (*recalibration of coarse models*). The paper describes the recalibration technique and then evaluates the accuracy of recalibrated models numerically on contrived examples using two techniques: u-plot and prequential likelihood, developed by others for software reliability growth models. The results indicate that the recalibrated predictions are often more accurate than the predictions obtained with the less detailed models, although this is not guaranteed. The techniques used to assess the accuracy of the predictions are accurate enough for one to be able to choose the model giving the most accurate prediction.

Keywords: Bayesian reliability assessment, fault tolerance, prediction accuracy.

1. Introduction

Off-the-shelf (OTS) components containing software are used widely for both development of new systems and upgrading existing (i.e. legacy) ones as part of their maintenance. The main reason for the trend is the low cost of OTS components compared with a bespoke development.

Dependability assessment of systems built with off-the-shelf components, however, is problematic [1]. The components are assessed typically in isolation or as part of their working in different operational environments. Porting the knowledge about their dependability acquired in one operational environment to a new operational environment is problematic. Unless software is free of design faults a good reliability record in the past in a different operational environment *does not guarantee* that it will also work reliably in the new operational environment, although the past record can be used to back *a priori* expectations for reliable operation in the new environment.

This difficulty in reusing assessment results has led some industries to look for solutions of assessment/certification without reference to the particular operational environment. For instance, the automotive industry introduced the concept of "Safety Element out of Context" (ISO 26262) to allow for certification of off-the-shelf components outside a specific operational environment. It remains to be seen how this idea will be used in practice – we hope "Safety Element out of Context" will be used as evidence of an *adequate process* and, thus, inform a *reasonable belief in good quality*, rather than as a substitute for assessment in the intended operational environment.

This paper applies Bayesian dependability assessment when a non black-box model of a system built with components is used. The paper uses several non black-box models which apply different degrees of abstraction to the system structure. Each of the non black-box models thus allows for different details to be accounted for which remain hidden by a black-box system model. There are several reasons to choose the non black-box over the black-box system model:

- Black-box model is not practicable for assessing ultra reliable software [2]. Indeed, using Bayesian assessment to get high confidence in ultra-high reliability will require an unrealistic amount of statistical testing unless one starts the assessment with ultra-strong prior belief. With non black-box model the confidence may grow faster [3].
- Even if black-box is adequate as a model for the intended new operational environment, one needs to build a prior distribution for system reliability, e.g. system probability of failure on demand (*pfd*) from the available evidence about reliability of the components used in the system. In this case one would ideally like to use all the evidence available in the prior– good reliability record about the part of the system which remains unchanged and a good record, possibly in a different operational environment, about the new components added to the system. Defining a prior distribution for the system reliability, which accounts for these sources of evidence, requires a non black-box model of the system.

Bayesian assessment with a non black-box model, however, poses its own difficulties:

- For any Bayesian inference defining an adequate prior is problematic. With a non black-box system model a *multivariate inference* is needed for which the usual difficulty of justifying the prior becomes even more difficult. Counterintuitive dependencies – how a change of my belief about X should influence my belief about Y – need to be quantified, which is really very hard [4].
- Using conjugate families [5] to alleviate the difficulty in defining multivariate distributions may have unpleasant consequences as reported in [3].
- Applying Bayesian inference numerically is a computationally intensive problem as it requires computing of multiple integrals over multivariate distributions of typically very small probabilities (e.g. the probabilities of failure of software components and of multiple software components failing together).

This paper's approach to the problem is pragmatic. System reliability is assessed by decomposing the system into *manageable views*. Each of the views focuses on a particular, typically small, part of the system (sub-system). The key element in the approach is that the views are defined in such a way that each sub-system appears in *multiple views*: e.g. one in which the particular sub-system is treated as a black-box (a *coarse view*) and one in which the same sub-system is shown in greater detail (a *refined view*). Bayesian inference can be applied with each of the views. Thus for a sub-system included in two views two predictions of this sub-system *pfd* will exist – from the coarse and from the refined views. The key element of the approach presented here is 'recalibrating' the coarse view using the predictions obtained with the refined view.

A central part in our approach is measuring the accuracy of the predictions obtained with the coarse and recalibrated view and choosing the one that is most accurate for the data at hand. For this we adapt techniques developed in software reliability growth modelling and discuss their suitability to our context.

The approach proposed in this paper can be applied to systems with different complexity, e.g. systems with structures ranging from a handful of components to those with a large number of components. The paper presents how the approach works using a system with a relatively simple structure: a legacy software system which is upgraded with a fault-tolerant component. The choice of such a system is dictated by several considerations: i) the example is important in practice - our previous consultancy work involved looking at exactly this system structure; ii) limiting system complexity allows for clear description of the proposed approach without having to deal with unnecessary system complexity. Scalability of the approach is discussed separately in the Discussion section.

The paper is arranged as follows. Section 2 states the problem addressed in the paper; Section 3 presents the approach of multi-views Bayesian assessment. Section 4 evaluates the usefulness of the approach using a series of contrived examples and Section 5 discusses the findings and their practical implications. Section 6 summarises the relevant research. Finally, Section 7 summarises the results and outlines directions for future work.

2. The problem

Consider the system schematically shown in Figure 1. It represents a legacy system subject to an upgrade: the rest of the system (ROS) is upgraded with a 1-out-of-2 fault-tolerant component (FT-component), which in turn consists of two components, channel A and channel B, which are integrated with the rest of the system via an interface. An example of an FT-component might be a reliable smart sensor in which two or more devices of similar functionality, but based on different physical principle for measurement are used. An upgrade of the FT-component is a typical practical problem for safety-critical applications with very long life. For such applications it is typical that long before they are decommissioned the components used in the original design are likely to cease to be manufactured by the respective vendors due to technological improvements. If the FT component has to be replaced the only viable option is for it to be upgraded with newer components.

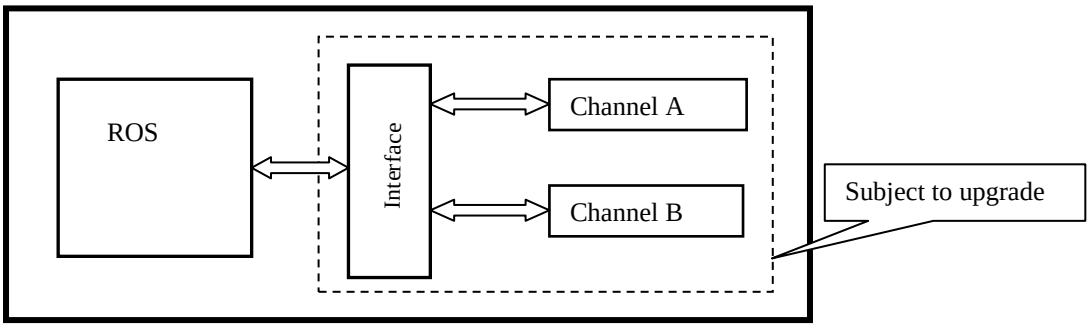

Figure 1. A 'legacy' system with a fault-tolerant component subject to an upgrade. The fault-tolerant component is a 1-out-of-2 system, the two channels of which are integrated with the 'rest of the system' (ROS) through an interface. As a result of the upgrade the entire sub-system – both channels and the interface – will be replaced. ROS remains unchanged.

Before we give a probabilistic description of the model presented in Figure 1 a few words about the 'components' of the system. ROS[1] and the two channels, A and B, are self-explanatory. The fourth box, 'Interface', encompasses

[1] ROS itself can consist of many individual components. This complexity, however, is ignored here. We initially limit the level of detail to keep the system structure simple and discuss the scalability aspects including the impact of the ROS's structure in the Discussion section.

both the tangible substance of making it possible for the two channels to work with the ROS, e.g. wiring and signal processing hardware, and also the design of the interaction such as protocols, constraints, which may be implemented in software. The 'Interface' may fail in different ways – as a result of a fault in either its tangible part or as a result of a design fault. In a sense when it comes to modelling the failures of the system the 'Interface' will take the 'blame' for all failures which cannot be attributed to the other three components.

2.1. Bayesian inference

A Bayesian approach to reliability assessment of an upgraded *on-demand system* is used in this paper. The *probability of failure on demand* (*pfd*) is the measure of interest. The *pfd* of any component and of the system itself are treated as *random variables* to express the epistemic uncertainty about the value of the respective *pfd*s. Some values of *pfd* are more believable than some other values and this is expressed by a probability distribution associated with the random variables.

Prior to observing the system in operation various sources of knowledge, e.g. previous experience with similar systems, are used to justify the use of a particular distribution as an adequate representation of uncertainty of the assessor about the *pfd* in question. When the system is put in operation its failures/successes in processing different demands provide new evidence about the reliability of the system and of the components thereof. As a result, the assessor's belief may change. Intuitively, frequent system failures should decrease the assessor's confidence in high reliability (i.e. low values of the corresponding *pfd* become less believable than higher values of the *pfd*) while very rare failures will increase confidence in high reliability.

If we treat the system as a *black box*, i.e. we can only distinguish between *system* failures and successes (Figure 2), the inference proceeds as follows. Denoting the system *pfd* as *p*, the posterior distribution of *p* after seeing *r* failures in *n* demands can be expressed as:

$$f_p(x \mid r,n) \propto L(n,r \mid x)\, f_p(x), \tag{1}$$

where $L(n,r \mid x)$ is the *likelihood* of observing *r* failures in *n* demands *if* the *pfd* were exactly *x*. This is given in this case (of independent demands) by the *binomial* distribution, $L(n,r \mid x) = \binom{n}{r} x^r (1-x)^{n-r}$. $f_p(\bullet)$ is the prior distribution of *p*, which represents the assessor's belief about *p*, *before* seeing the results of how the system processes *n* demands.

The paper uses the convention of *upper case letters* to denote a random variable, e.g. P_A, while lower case letters are used to refer to the values that a random variable takes, i.e. p_A

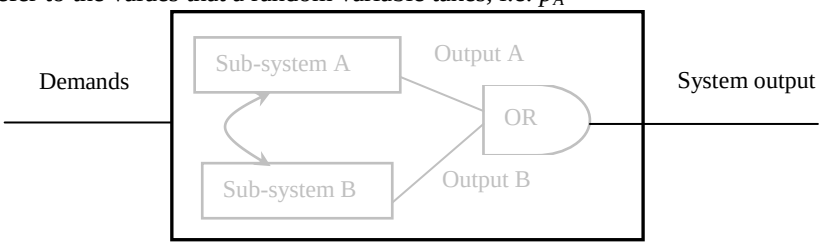

Figure 2. Black-box model of a system. The internal structure of the system is unknown and cannot be used in the inference. Only the system scores on each demand – success or failure – are recorded on each demand and are fed into the inference procedure.

With the non black-box model the scores at the *outputs of the sub-systems* (i.e. 0 - success or 1 - failure) are explicitly taken into account. They are recorded for every demand processed by the system and are then used in the inference. The system score can be a *deterministic* or a *non-deterministic function* of the scores of the sub-systems. If the system score is a deterministic function (called *structure function*, [6]) the system *pfd* can be expressed using the *pfd* of the components. Typical examples are serial and parallel systems, but the structure function of systems with more components may be significantly more complex.

If the system failure is a non-deterministic function of the components, the system score needs to be modelled explicitly.

A more refined analysis with the non black-box model would require knowledge of how many times each of the sub-systems is executed per demand (the control flow of executing a demand). This paper limits analysis to assuming that *all sub-systems are used exactly once per demand*.

The example of an upgraded system, Figure 1, has a vector of component scores of four components, $S_{ROS}S_{Interface}S_{ChA}S_{ChB}$, and, hence, 16 different score combinations (0000, 0001, …, 1111) can be observed on a demand. 0000, for instance, represents the demands on which none of the components fails; 0001 represents the demands on which ROS, the Interface, and channel A, do not fail while channel B fails, etc. Finally, 1111 represents the demands on which all four sub-systems coincidentally fail. Each of these score combinations can be observed on a randomly chosen demand.

Probabilities: p_{0000}, p_{0001}, …, p_{1111} are associated respectively with each of the possible outcomes. Clearly, these probabilities sum up to 1 (with certainty one of the score vectors will be observed when the system is called upon), i.e.:

$$p_{0000}+p_{0001}+p_{0010}+p_{0011}+p_{0100}+p_{0101}+p_{0110}+p_{0111}+p_{1000}+$$
$$p_{1001}+p_{1010}+p_{1011}+p_{1100}+p_{1101}+p_{1110}+p_{1111}=1$$

In other words, there are 15 degrees of freedom in describing the joint behaviour of the system. The probabilities, p_{0000}, p_{0001}, …, p_{1111}, are clearly related to the *pfd* of the sub-systems. For instance, the probability of failure of ROS, p_{ROS}, is the sum of the probabilities p_{1XXX}, where 'X' can be either 0 or 1, i.e. when ROS has failed no matter whether the other sub-systems failed or not on the same demand. Formally, this can be expressed as follows:

$P_{ROS} = p_{1000} + p_{1001} + p_{1010} + p_{1011} + p_{1100} + p_{1101} + p_{1110} + p_{1111}$.

Similar expressions can be derived for the marginal probabilities of failure of the other components: $p_{Interface}$, and the two channels of the fault-tolerant sub-system $p_{ChannelA}$ and $p_{ChannelB}$.

Non black-box Bayesian inference in the system (Figure 1) will, thus, require a 15-variate prior distribution to be defined, which the Bayesian inference will transform to a *posterior* distribution using the prior and the score vectors observed in operation. The variates of the joint distribution can be any 15 out of the 16 probabilities listed above, p_{0000}, p_{0001}, …, p_{1111}, or any function defined on these probabilities, e.g. some of the marginal probabilities of the components. The joint distribution will describe the assessor's belief/knowledge about the system's failure behaviour prior to and after the observations, respectively.

Various distributions of interest can be derived from the joint posterior distribution, e.g. the *pfd* of any of the components, subsystem or system *pfd* in case these are deterministic functions of the *pfds* of the components. If the *pfd* of interest is explicitly a variate in the used posterior distribution, deriving the marginal distribution of the *pfd* merely requires integrating out the nuisance parameters of the posterior distribution. If the *pfd* of interest is not explicitly represented in the joint distribution, then a transformation is required of the prior/posterior joint distribution to another distribution in which the *pfd* of interest appears explicitly as one of the variates of the distribution. Such a transformation is a standard operation from calculus. From the transformed joint distribution the marginal distribution of interest will be derived by integrating out the *nuisance* parameters.

If system failure is a non-deterministic function of the sub-system scores, we will need to add an extra variate – the system score. For our problem this addition will require a 31-variate distribution in order to describe probabilistically the failure behaviour of the system.

Clearly the full model of the system, Figure 1, is very complex. Even if the system score is a deterministic function of the component scores, one needs (at least) a 15-variate distribution and to do inference with it! The problems associated with a multivariate Bayesian inference (difficulty to define sensibly multivatiate priors and computational difficulties with the inference itself) make a good case for an effort to simplify the multivariate distribution used in the inference. Using a simplified inference, however creates a new problem – controlling the error due to the simplifications introduced. In both cases there will be errors: in the former case this will be the error due to poor prior while in the latter case – the error due to the model simplification made. It is far from obvious which is better – using a full inference starting with a poor prior, i.e. one which captures poorly the true belief of the assessors or using a simplified prior which defines the problem more simply, requires fewer parameters from the assessor and thence may capture more accurately the true belief of the assessor, but ignores some aspects of the system model, which may turn out to be important.

2.2. Simplifying the prior distributions

An extreme case of simplifying the model of the system is treating the system as a black-box. This simplification can be applied to any system. In this case only a univariate distribution of the system *pfd* is required and the inference is computationally trivial but, as discussed above, is problematic for ultra-high reliability ranges [2] and in integrating in the prior the previous knowledge that might be available about the reliability of components.

If the system is modelled as a non black-box several options exist for simplifying the system model, which in turn will reduce the complexity of the prior and of the inference.

- Assume that some of the components are *perfect*. In the case of an upgrade (Figure 1), for instance, assuming that the *interface* between ROS and the upgraded components *is perfect* (i.e. does not contain design faults) may be relatively easy to justify, e.g. if the integration is 'trivial' or can be tested exhaustively. Such a simplification in the example system will reduce the size of the distribution needed to a 7-variate distribution. Although this is a significant reduction in comparison with the 15-variate case, it still poses difficulties for the assessor to specify a sensible prior.
- Assume some form of independence between the variates in the joint prior distribution. Assuming *independence between the uncertainties* associated with the variates in the prior, for instance, one might be able to justify assuming independence between the uncertainties associated with the *pfd* of ROS and of the fault-tolerant component by lack of evidence on the contrary. Note that such an assumption is different from assuming that the components fail independently, an assumption for reliability of hardware components [7]. In

the former case we merely assume that the *pfd*s of the ROS and of the FT-component are independently distributed random variables, while in the latter case we postulate that whatever the 'true' probabilities of failure of the individual channels the probability of their failing together will be a product of these 'true' probabilities. In other words in the latter case we postulate that we know *with certainty* how to derive the probability of simultaneous failure from the marginal probabilities of failure of ROS and of the FT-component and that the only uncertainty in the model is associated with those marginal probabilities. The two assumptions of independence have very different implications. In the former case (independent distributions) an erroneous assumption can be 'fixed' when the system is observed in operation and the collected data suggest that the components tend to fail non-independently. On the contrary, once the failures of the components are assumed to be independent, even very strong evidence against the assumption cannot fix the error.

- The model can also be simplified by making various *conditional independence* assumptions. This alternative line of reasoning is not pursued in this paper but is discussed briefly in Section 5.
- Simplify the inference by decomposing the system model into simpler 'views', each of which is tractable both computationally and in terms of defining plausible priors. The example system from Figure 1 can be decomposed as shown in Figure 3 using two 'views', the *course view* of the system and the *refined view* of the FT components. Each of these models consists of two components and their probabilistic description (3-variate joint distributions) can be *derived* from the full multivariate distribution by integrating out *different nuisance parameters*. The two 3-variate models are *significantly simpler* than the full non black-box system model. Each of the models can be used on its own for Bayesian assessment of the sub-system it represents. A key element of the paper proposes a way of combining the results of the two inferences, by *recalibrating* the posterior derived with the coarse view using the posterior derived with the refined view. In order to be able to do that we require that the views be chosen to allow for *alternative ways* of expressing the *same variate* – in this example the *pfd* of the FT-component.

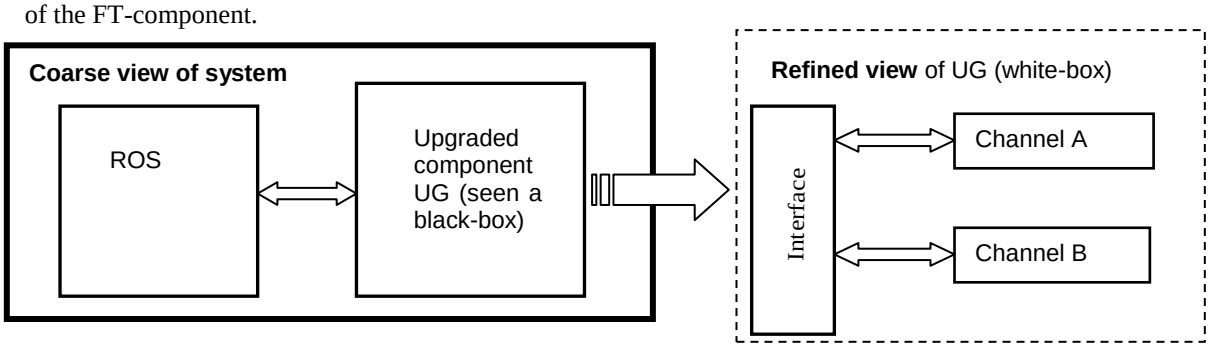

Figure 3. Two-tier view on the system: In the 'coarse view' the FT-component is represented as a black-box. In the 'refined view' this black-box is detailed as a 1-out-of-2 system. In the 'coarse' view the system consists of two components – ROS, and the FT-component. The refined view consists of the channels of the FT-component only.

Note that the coarse view is more detailed than a black-box view of the system. The coarse view allows us to use the knowledge accumulated to date about the quality of ROS when specifying the prior for this view, while if the black-box view is used this is not explicitly required, instead the black-box view only requires a distribution of the system *pfd* to be specified, which will clearly be affected by the *pfd* of ROS, but in a complex way. The refined view, on the other hand, allows us to use the detailed information about the two channels of the fault-tolerant component and distinguish between the four possible observations of these two channels in operation (each may fail or succeed on a demand); such detailed knowledge is ignored in the coarse view. In summary, although each of the two views defined above differs from the full model of the system, they account for important aspects of the components of the system.

Decomposing the system model into partial views replaces the problematic inference using the full system model with new problems:

- how to combine the two posteriors derived with the two partial models (views), and
- finding out if combining the two views improves the prediction accuracy in comparison with the simpler alternatives such as using the coarse view or the black-box model of the system, which offer ways of expressing the system *pfd*.

Assessing the predictive accuracy of any inference is a problem relatively separate from the inference itself. Prediction accuracy has been studied in *software reliability growth modelling*. Techniques for assessing prediction accuracy *objectively* have been developed such as u-plot and prequential likelihood [8], which are applied here to the problem.

Note at this point that we may construct three different prediction models of system *pfd*:

- the black-box model, ignoring entirely any knowledge about the internal structure of the systems;
- the coarse view (ignoring details about the FT-component structure);
- recalibrated model which uses the predictions of the *pfd* of the FT-component to recalibrate the coarse model and obtain a prediction of the system *pfd*.

3. The solution: a multi-view Bayesian inference

This section describes the main contribution of the paper, a multi-view Bayesian inferences procedure.
The procedure consists of the following steps:

- **Step 1:** Bayesian inference using a simplified system model (a "coarse" system model) in which the FT-component is treated as a black-box;
- **Step 2:** Bayesian inference with the FT-component only, in which the behaviour of the channels is modelled explicitly (a non black-box model of the FT-component is used) but the rest of the system is ignored;
- **Step 3:** Recalibration of the system posterior is derived using the posteriors derived in Step 1 and step 2 above.

3.1. The coarse view system model: the FT-component viewed as a black-box

Sub-systems ROS and UG (Figure 3) are assumed imperfect and their probabilities of failure are assumed uncertain, quantified using a probability distribution. We further assume that the 'Interface' component (Figure 1) is perfect, i.e. does not contain faults (design or any other). The scores of the sub-systems, which can be observed on a randomly chosen demand, are summarised in Table 1 (1 is used when the component fails to process the demand correctly, 0 – when the demand is processed correctly by the component).
A 3-variate joint distribution is defined in which the variates can be any three out of the four probabilities listed in column three of Table 1, e.g. $f_{p_{01},p_{10},p_{11}}(\bullet,\bullet,\bullet)$, or derived from them.

Given a set of observations, r_1, r_2, and r_3 in n demands, the posterior distribution will be:

$$f_{p_{01},p_{10},p_{11}}(x,y,z \mid n,r_1,r_2,r_3) = \frac{f_{p_{01},p_{10},p_{11}}(x,y,z)L(n,r_1,r_2,r_3 \mid p_{01},p_{10},p_{11})}{\iiint\limits_{p_{01},p_{10},p_{11}} f_{p_{01},p_{10},p_{11}}(x,y,z)L(n,r_1,r_2,r_3 \mid p_{01},p_{10},p_{11})dxdydz} \tag{2}$$

where

$$L(n,r_1,r_2,r_3 \mid p_{01},p_{10},p_{11}) = \frac{n!}{r_1!r_2!r_3!(n-r_1-r_2-r_3)!} {p_{01}}^{r_1} {p_{10}}^{r_2} {p_{11}}^{r_3} (1-p_{01}-p_{10}-p_{11})^{n-r_1-r_2-r_3} \tag{3}$$

is the multinomial likelihood of the observation (r_1, r_2, and r_3 in n demands).

Table 1

Component scores			
ROS	UG	Probability	Observations in n demands
0	0	p_{00}	r_0
0	1	p_{01}	r_1
1	0	p_{02}	r_2
1	1	p_{11}	r_3

Table 1**.** The combinations of component scores, which can be observed on a randomly chosen demand, are shown in the first two columns. The notations, used for the probabilities of these combinations on a randomly chosen demand, are shown in the third column. Finally, the number of times the various score combinations are observed in n trials ($n = r_0 + r_1 + r_2 + r_3$), are shown in the last column of the table. Note that r_2 represents the number of failures of the fault-tolerant component UG, i.e. when both channels of this FT-component fail simultaneously. The individual channel failures are tolerated and not counted in this view.

It is intuitively easier to express the prior in terms of probability of failure of the components ROS and UG. Clearly, the probabilities of failure of these, p_{ROS} and p_{UG}, respectively, can be expressed as:
$p_{ROS} = p_{10} + p_{11}$ and $p_{UG} = p_{01} + p_{11}$.
$p_{ROS \wedge UG}$ represents the probability of simultaneous failure of both ROS and UG on the same demand, hence, $p_{11} \equiv p_{ROS \wedge UG}$. The prior distribution $f_{p_{01},p_{10},p_{11}}(\bullet,\bullet,\bullet)$ can be transformed to a new set of variates, p_{ROS}, p_{UG}, and $p_{ROS \wedge UG}$, $f_{p_{ROS},p_{UG},p_{ROS \wedge UG}}(\bullet,\bullet,\bullet)$, which can be easier for an assessor to work with[2]. Given the same observation, (r_1, r_2, and r_3 in n demands), it can be shown that the posterior distribution can be expressed as:

$$f_{p_{ROS},p_{UG},p_{ROS \wedge UG}}(x,y,z \mid n,r_1,r_2,r_3) = \frac{f_{P_{ROS},P_{UG},P_{ROS,UG}}(x,y,z)L(n,r_1,r_2,r_3 \mid p_{ROS},p_{UG},p_{ROS,UG})}{\iiint\limits_{p_A,p_B,p_{AB}} f_{p_{ROS},p_{UG},p_{ROS,UG}}(x,y,z)L(n,r_1,r_2,r_3 \mid p_{ROS},p_{UG},p_{ROS,UG})dxdydz} \tag{4}$$

where

[2] The transformation (change of variable) is a standard procedure in calculus which requires the Jacobian to be computed for the old and the new set of variates. For the particular transformation dealt with here, the Jacobian is equal to 1.

$$L(n, r_1, r_2, r_3 \mid p_{ROS}, p_{UG}, p_{ROS \wedge UG}) =$$

$$\frac{n!}{r_1! r_2! r_3! (N - r_1 - r_2 - r_3)!} (p_{ROS} - p_{ROS \wedge UG})^{r_2} (p_{UG} - p_{ROS \wedge UG})^{r_1} p_{ROS \wedge UG}^{\ r_3} (1 - p_{ROS} - p_{UG} + p_{ROS \wedge UG})^{n - r_1 - r_2 - r_3}$$

is the likelihood of the observation for the new set of variates used in the prior/posterior distributions.

Up to this point the inference will be the same no matter how the event 'system failure' is defined. The marginal distribution of the *system pfd*, however, is affected by the structure model, i.e. how system failure is defined. In this paper we consider a serial system of ROS and UG (Figure 1): a failure of either of these two sub-systems leads to a system failure. Deriving the distribution of system failure would require a transformation of the posterior distribution, $f_{p_{ROS}, p_{UG}, p_{ROS \wedge UG}}(\bullet, \bullet, \bullet \mid n, r_1, r_2, r_3)$, to a new distribution, $f_{p_{ROS}, p_{UG}, p_S}(\bullet, \bullet, \bullet \mid n, r_1, r_2, r_3)$, where for any values of P_{ROS}, P_{UG} and $P_{ROS \wedge UG}$ the new variable P_S is defined as: $P_S = P_{ROS} + P_{UG} - P_{ROS \wedge UG}$. From the transformed joint distribution $f_{p_{ROS}, p_{UG}, p_S}(\bullet, \bullet, \bullet \mid n, r_1, r_2, r_3)$ the marginal distribution of P_S, $f_{p_S}(\bullet \mid n, r_1, r_2, r_3)$, will be derived from $f_{p_{ROS}, p_{UG}, p_S}(\bullet, \bullet, \bullet \mid n, r_1, r_2, r_3)$ by integrating out the nuisance parameters P_{ROS} and P_{UG}.

3.2. The refined view model: a detailed model of the upgraded FT-component

This model is practically identical to the coarse model, the only difference being the notations used.

The channels A and B of the FT-component are both assumed imperfect and their probabilities of failure are assumed uncertain, the uncertainty being quantified by a probability distribution, which is detailed below. The scores of the channels, which can be observed on a randomly chosen demand, are summarised in Table 2.

A 3-variate joint distribution must be defined in which any three of the probabilities shown in Table 2 can be used as variates. Let us choose to use the probability distribution $f_{p_{\bar{A}\bar{B}}, p_{A\bar{B}}, p_{AB}}(\bullet, \bullet, \bullet)$.

Table 2

Channel scores			
Channel A	Channel B	Probability	Observations in n demands
0	0	$p_{\bar{A}\bar{B}}$	r_4
0	1	$p_{\bar{A}B}$	r_5
1	0	$p_{A\bar{B}}$	r_6
1	1	p_{AB}	r_7

Table 2. The combinations of channel scores, which can be observed on a randomly chosen demand, are shown in columns one and two. The notations used for the probabilities of these combinations are shown in column three while the number of times the score combinations are observed in *n* trials, $r_4 - r_7$, respectively ($n = r_4 + r_5 + r_6 + r_7$), are shown in the last column of the table.

Given a set of observations, r_5, r_6, and r_7 in n demands, the posterior distribution will be:

$$f_{p_{\bar{A}\bar{B}}, p_{A\bar{B}}, p_{AB}}(x, y, z \mid n, r_5, r_6, r_7) = \frac{f_{p_{\bar{A}\bar{B}}, p_{A\bar{B}}, p_{AB}}(x, y, z) L(n, r_5, r_6, r_7 \mid p_{\bar{A}\bar{B}}, p_{A\bar{B}}, p_{AB})}{\iiint_{p_{\bar{A}\bar{B}}, p_{A\bar{B}}, p_{AB}} f_{p_{\bar{A}\bar{B}}, p_{A\bar{B}}, p_{AB}}(x, y, z) L(n, r_5, r_6, r_7 \mid p_{\bar{A}\bar{B}}, p_{A\bar{B}}, p_{AB}) dx dy dz} \quad (5)$$

where

$$L(n, r_5, r_6, r_7 \mid p_{01}, p_{10}, p_{11}) =$$

$$\frac{n!}{r_5! r_6! r_7! (n - r_5 - r_6 - r_7)!} p_{\bar{A}B}^{\ r_5} p_{A\bar{B}}^{\ r_6} p_{AB}^{\ r_7} \left(1 - p_{\bar{A}B} - p_{A\bar{B}} - p_{AB}\right)^{n - r_5 - r_6 - r_7} \quad (6)$$

is the multinomial likelihood of the observation (r_5, r_6, and r_7 in n demands).

It may be intuitively easier to express the prior in terms of the probability of failure of the channels, A and B, which can be expressed as follows:

$p_A = p_{A\bar{B}} + p_{AB}$ and $p_B = p_{\bar{A}B} + p_{AB}$.

P_{AB} represents the probability of coincident failures of both channels, A and B, on the same demand. The prior distribution $f_{p_{\bar{A}\bar{B}}, p_{A\bar{B}}, p_{AB}}(\bullet, \bullet, \bullet)$ can be transformed to a new set of variates, P_A, P_B, and P_{AB}, $f_{p_A, p_B, p_{AB}}(\bullet, \bullet, \bullet)$, which can be easier for an assessor to work with.

Given the same observation, (r_5, r_6, and r_7 in n demands), it can be shown that the posterior distribution can be calculated as:

$$f_{p_A,p_B,p_{AB}}(x,y,z \mid n,r_5,r_6,r_7) = \frac{f_{p_A,p_B,p_{AB}}(x,y,z)L(n,r_5,r_6,r_7 \mid p_A,p_B,p_{AB})}{\iiint\limits_{p_A,p_B,p_{AB}} f_{p_A,p_B,p_{AB}}(x,y,z)L(n,r_5,r_6,r_7 \mid p_A,p_B,p_{AB})dxdydz} \quad (7)$$

where

$$L(n,r_5,r_6,r_7 \mid p_A,p_B,p_{AB}) = \frac{n!}{r_5!r_6!r_7!(n-r_5-r_6-r_7)!}(p_A-p_{AB})^{r_6}(p_B-p_{AB})^{r_5}\,p_{AB}{}^{r_7}(1-p_A-p_B+p_{AB})^{n-r_5-r_6-r_7}$$

is the likelihood of the observation for the new set of variates used in the prior/posterior distributions.

In this model we are only interested in the probability of failure of the FT component, i.e. in P_{AB}, which can be derived from the posterior distribution (7) after integrating out the nuisance parameters, P_A and P_B.

3.3. Recalibrated system model: combining the coarse and the refined views

Now the procedure of *recalibrating the* posterior obtained with the *coarse view* using the posterior from the refined view on the FT component is described.

One of the variates of the distribution, $f_{p_{ROS},p_{UG},p_{ROS\wedge UG}}(x,y,z \mid n,r_1,r_2,r_3)$, used with the coarse view model (4), is P_{UG}. In the refined view the same *pfd*, P_{AB}, is expressed differently (7). Even if the coarse and refined views are used with consistent marginal priors, e.g. $f_{p_{AB}}(\bullet) = f_{p_{UG}}(\bullet)$, due to using different information in both inferences the posterior distributions, in general, will be different, i.e. $f_{p_{AB}}(\bullet \mid n,r_5,r_6,r_7) \neq f_{p_{UG}}(\bullet \mid n,r_1,r_2,r_3)$. The reader should have noticed by now that $r_7 = r_1 + r_3$, i.e. the total number of failures of the FT-component (i.e. simultaneous failures of channel A and channel B in the refined view) is equal to the number of observations when the UG component fails on its own or simultaneously with ROS.

In the coarse view, Section 3.1, by observing the system in operation we have learned about the epistemic uncertainties about the probabilities of failure of ROS and UG (including the *dependencies* between the variates). In the refined view this knowledge is not acquired. Instead, the refined view concentrates on learning how effective fault tolerance is for the FT-component, but ignores how it relates to ROS, i.e. whether there is evidence that ROS and the FT-component are failing independently or their failures are correlated – positively or negatively.

The recalibration proposed here is based on the *following rationale*. We assume that the *coarse system model captures accurately the epistemic uncertainty about the pfds of ROS and UG* and their dependence. Formally, this is represented by the conditional distribution:

$$f_{p_{ROS},p_{ROS\wedge UG}|p_{UG}}(\bullet,\bullet \mid P_{UG}=z) = \frac{f_{p_{ROS},p_{UG},p_{ROS\wedge UG}}(\bullet,\bullet,z)}{f_{p_{UG}}(z)} \quad (8)$$

We postulate that the conditional distribution (8) remains unaffected by the recalibration.

In other words, we postulate that the recalibrated posterior, $f_{p_{ROS},p_{AB},p_{ROS\wedge UG}}(\bullet,\bullet,z)$, is defined as:

$$f_{p_{ROS},p_{AB},p_{ROS\wedge AB}}(\bullet,\bullet,z) = f_{p_{ROS},p_{ROS\wedge UG}|p_{UG}}(\bullet,\bullet \mid P_{UG}=z)\,f_{p_{AB}}(z) = \frac{f_{p_{ROS},p_{UG},p_{ROS\wedge UG}}(\bullet,\bullet,z)}{f_{p_{UG}}(z)}\,f_{p_{AB}}(z). \quad (9)$$

4. Empirical Evaluation

The inferences with the coarse and the recalibrated models differ significantly in *complexity*. Recalibration requires inferences with both the coarse and the refined view models and in addition a transformation of the entire posterior distribution obtained with the coarse model as suggested by equation (9).

The obvious question is whether the extra complexity of recalibrated inference brings advantages in comparison with the simpler inferences. If it does, are they guaranteed whatever the priors and the observations or is more detailed analysis needed to identify whether recalibrated inference brings advantages with the specific problem at hand.

The essential question is whether we can trust the predictions obtained with coarse and the recalibrated models. If the predictions they produce are the same, can we assume that they are accurate or despite their being similar do we still have to validate their accuracy? If, instead, the predictions obtained with the two models are different can we trust any of them and if so which one? The answers to these questions are not obvious. Clearly, both views ignore some aspects of system behaviour and thus introduce errors. We would like to scrutinise various aspects of these errors and come up with a practical advice on deciding if any of the predictions are accurate or not.

This section starts with a contrived example to illustrate that the predictions obtained with the recalibrated inference may differ significantly from the predictions obtained with coarse system model. It then proceeds with a more detailed analysis of accuracy using a series of carefully chosen contrived numerical examples.

4.1. An example

Consider an example parameterisation of the system, which uses the coarse and the recalibrated system models identified above. We define the prior distributions needed as follows.

- The **FT-component**. The prior, $f_{p_A,p_B,p_{AB}}(\bullet,\bullet,\bullet)$, is constructed under the assumption that $f_{p_A}(\bullet)$ and $f_{p_B}(\bullet)$ are both Beta distributions, $B(\bullet,\alpha,\beta)$, in the interval [0, 0.01] and are independent of each other, i.e. $f_{p_A,p_B}(\bullet,\bullet) = f_{p_A}(\bullet) f_{p_B}(\bullet)$. The parameters α and β used in the examples for the two distributions are: α_A = 2, β_A = 2 for channel A and α_B =3, β_B = 3 for channel B, respectively. The conditional distributions, $f_{p_{AB}|p_B,p_A}(\bullet \mid P_A = p_A, P_B = p_B)$, for every pair of values of P_A and P_B, are defined as Beta distributions, $B(\bullet,a,b)$ in the range [0, min(p_A, p_B)] with parameters α_{AB} = 2, β_{AB} = 2. These conditional distributions together with the defined above joint distributions $f_{p_A,p_B}(\bullet,\bullet)$ completely define the tri-variate prior distribution, $f_{p_A,p_B,p_{AB}}(\bullet,\bullet,\bullet)$, used in the refined view of the FT-component. The marginal distribution $f_{p_{AB}}(\bullet)$ is derived from $f_{p_A,p_B,p_{AB}}(\bullet,\bullet,\bullet)$ and used in the prior of the coarse view. No claims are made that the priors used in the examples should be used for practical assessment. They serve illustrative purposes only and yet, have been chosen from a reasonable range. Each of the channels, for instance, has an average *pfd* of 5.10^{-3}, which is a value from a typical range for many applications.
- The **coarse system view**. The system prior, $f_{p_{ROS},p_{UG},p_{ROS\wedge UG}}(\bullet,\bullet,\bullet)$, is constructed under the assumption that the *pfd* of the ROS and of the UG component are *independently distributed*, i.e. $f_{p_{ROS},p_{UG}}(\bullet,\bullet) = f_{p_{ROS}}(\bullet) f_{p_{UG}}(\bullet)$. $f_{p_{ROS}}(\bullet)$ is assumed to be a Beta distributions, $B(\bullet,\alpha,\beta)$, in the interval [0, 0.01] with parameters α_{ROS} = 3, β_{ROS} = 3, while $f_{p_{UG}}(\bullet) \equiv f_{p_{AB}}(\bullet)$ is as derived from the refined view of the FT-component. The conditional distributions, $f_{p_{ROS\wedge UG}|p_{ROS},p_{UG}}(\bullet \mid P_{ROS} = p_{ROS}, P_{UG} = p_{UG})$ for every pair of values of P_{ROS} and P_{UG}, are defined as Beta distributions, $B(\bullet,\alpha,\beta)$ in the range [0, min(p_{ROS}, p_{UG})] with parameters $\alpha_{ROS\wedge UG}$ = 2, $\beta_{ROS\wedge UG}$ = 2.
- The **observations**. n = 4000 trials, r_{ROS} = 4, $r_A = r_B$ = 10, r_{AB} = 2; $r_{ROS\wedge UG}$ = 2. In other words, the failures of the FT-component coincide with failures of theROS. One should also notice that 20% of the channel failures are simultaneous.

This example represents a complex relationship between the failures of the channels of the FT-component on the one hand and the failures of ROS and of the FT component on the other hand.

Figure 4 shows the difference in the predicted *pfd* of the FT-component obtained with the coarse system model and the refined model of the FT-component, while Figure 5 shows the system *pfd* obtained with the coarse and the recalibrated system models.

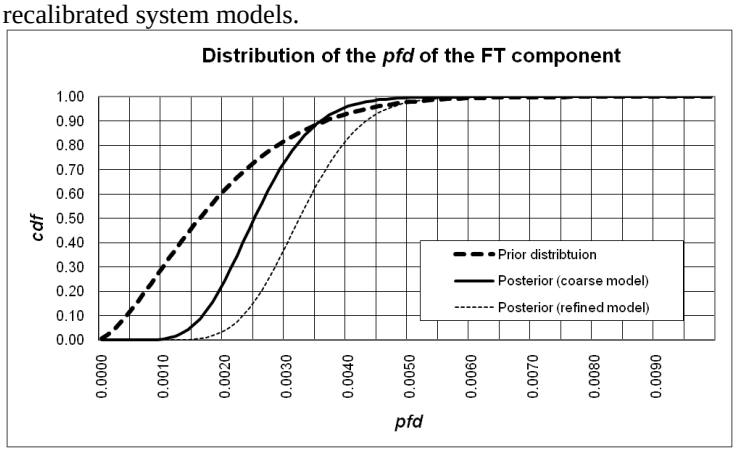

Figure 4. The FT-component *pfd* derived with the coarse and the refined models. The *pfd* derived with the refined model is significantly worse than the posterior derived with the coarse model. Both are significantly worse than the prior.

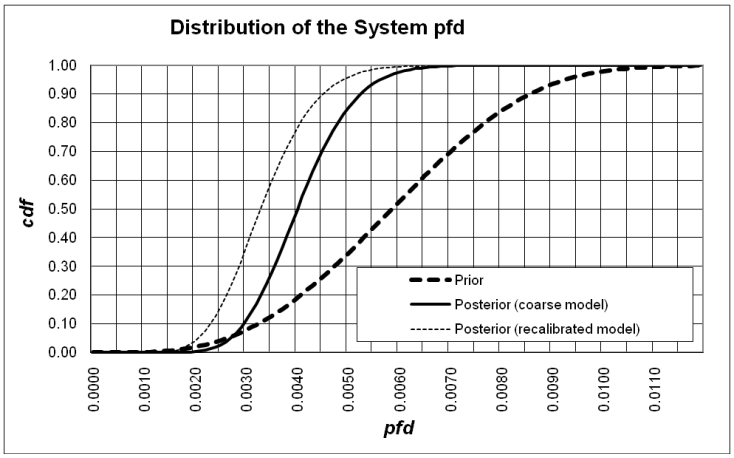

Figure 5. The posterior system *pfd* when ROS and the UG component form a serial system. The posterior obtained with the coarse model is significantly worse than the predictions obtained with the recalibrated model. Both clearly indicate that the system is better than was assumed in the prior.

The disagreement between the coarse and the refined views on the posterior *pfd* of the FT-component is significant (e.g. the confidence associated with a target *pfd* = 0.003 is 40% in the case of refined mode and almost 70+% with the coarse model). The refined posterior is clearly significantly more pessimistic than the one derived with the coarse model.

The disagreement between the coarse and the refined views on the posterior *system pfd* is significant, too, (e.g. the confidence associated with a system target *pfd* = 0.004 is 50% with the coarse model and almost 80% with the recalibrated model). The recalibrated posterior is clearly significantly more optimistic than the one derived with the coarse model. This outcome is very intriguing but not easy to explain. The coarse model has captured the strong positive correlation between the UG (the FT-component) and the ROS – all observed failures of the two happen to be simultaneous. But the system is a serial one and for it positive correlation between the failures of the components means that their failure regions (on the demand space) overlap significantly, thus reducing the probability of system failure for the same probability of failure of the components. The recalibration of the *system pfd* posterior introduced by (9) preserves the conditional distributions, $f_{p_{ROS}, p_{ROS \wedge UG} | p_{UG}}(\bullet,\bullet \mid P_{UG} = z)$ but changes the weights of the conditional distributions for the different values of the *pfd* of the FT-component. In the example, the posterior distribution of the FT-component obtained with the coarse model is *stochastically more optimistic* than the posterior obtained with the refined model of the FT-component. It appears that the recalibration leads to a better system pfd, but the mechanisms leading to an improved system *pfd* are not obvious.

4.2. Accuracy of Bayesian predictions

The difference between the predictions obtained with and without recalibration is not surprising, as already discussed. To these two prediction models we can, of course, add a *black-box* system model, which operates with even less detailed data than the coarse model does.

Littlewood et al. developed techniques for assessment of prediction accuracy of various *software reliability growth models* [8]. These methods allow one to compare *objectively* how two prediction models compare in terms of prediction accuracy (i.e. which of them gives systematically more accurate predictions) and even, under certain conditions, to assess whether the predictions are optimistic or pessimistic. The two techniques, *prequential likelihood* and the *u-plot*, are used to assess the accuracy of predictions obtained with the available models - the black-box, the coarse and the recalibrated models. Appendix 1 provides a summary of the techniques.

4.3. More examples

We now study the prediction accuracy of the models using data generated via *Monte-Carlo* simulation of the behaviour of the system under consideration (Figure 1). Four different systems are simulated using parameters as shown in Table 3. These are the *true values of the probabilities* that we try to estimate using Bayesian inference.

Table 3

	$System_1$	$System_2$	$System_3$	$System_4$
P_{ROS}	0	0	0.001	0.001
P_A|ROS failed	0	0	0.5	0.5
P_A|ROS did not fail	0.005	0.005	0.001	0.002
P_B|A did not fail, ROS did not fail	0	0.0035	0.005	0.002
P_B|A did not fail, ROS failed	0	0	0.5	0.002
P_B|A failed, ROS did not fail	1	0.3	0.005	0.002
P_B|A, ROS both failed	0	0	0.01	1.0
Number of demands simulated	50,000	50,000	100,000	50,000
P_{ROS} marginal	0	0	0.001	0.001

P_A marginal	0.005	0.005	0.0015	0.0025
P_B marginal	0.005	0.00498	0.00525	0.0025

The choice of parameters was motivated by trying to cover a *range of plausible scenarios*. The difference between $System_1$ – $System_4$ is in the reliability of individual components (the marginal probabilities of failure are shown at the bottom of the table) and in the level of correlation between the failures of the channels of the FT-component and between the failures of the FT-component and ROS. For $System_1$ the channels of the FT-component always fail simultaneously, and never together with ROS. $System_2$ is similar to $System_1$ in the sense that ROS and the channels of the FT-component never fail together, but there is only 30% chance that the channels of the FT-component fail simultaneously. With $System_3$ and $System_4$ there are chances of simultaneous failure between ROS and the channels of the FT-component. The difference between the two is in the degree of correlation between the failures of ROS on the one hand and of the channels of the FT-components on the other hand: with $System_4$ the correlation is much stronger than with $System_3$. Note that with $System_4$ channel B fails with certainty if both ROS and channel A have failed, while with $System_3$ in similar circumstances the probability of failure of channel B is only 1%.

In addition to the 'observations' collected by simulation the impact of the prior on the predictions was studied by defining different priors, $f_{p_{ROS},p_{UG},p_{ROS\wedge UG}}(\bullet,\bullet,\bullet)$, and $f_{p_A,p_B,p_{AB}}(\bullet,\bullet,\bullet)$, for the coarse model of the system and the refined model of the FT-component, respectively, which were constructed by making assumptions similar to those in the example given at the beginning of this section using a set of parameters: α_A, β_A, α_B, β_B, α_{AB}, β_{AB}. α_A, β_A for channel A, α_B, β_B for channel B, respectively, describe the uncertainty associated with the *pfd*s of the two channels. α_{AB}, β_{AB} characterise the uncertainty about the conditional probability of simultaneous failure of channel A and B. Similarly, α_{ROS}, β_{ROS}, $\alpha_{ROS\wedge UG}$, $\beta_{ROS\wedge UG}$, characterise the uncertainty about the probability of failure of ROS and the simultaneous failure of ROS and the FT-component, respectively. The parameters used in the study are shown in Table 4.

Table 4

	$Prior_1$	$Prior_2$	$Prior_3$	$Prior_4$
α_A	2	20	1	10
β_A	2	20	10	10
α_B	3	20	1	10
β_B	3	20	10	12
α_{AB}	1	1	1	1
β_{AB}	1	1	1	3
α_{ROS}	3	1	1	1
β_{ROS}	10	10	10	200
$\alpha_{ROS\wedge UG}$	3	3	1	1
$\beta_{ROS\wedge UG}$	3	3	10	10

Priors were chosen so that the degree of uncertainty in the *pfd*s of the two channels of the FT-component, A and B and the uncertainty in the *pfd*s of ROS and the failure correlation between ROS and UG also varied to enable study of how the prior affects the predictions.

The findings are now illustrated, grouping the observations as summarised in Table 5 below. In the first four examples inference was applied using the first 5,000 observations of the simulated 50,000 demands of $System_1$ and $System_2$, respectively, using a fixed testing campaign size of 100 demands. In the other two examples the size of the 'testing campaigns' varied between 25 and 200 showing the results with 200 campaigns, thus using between 5,000 and 40,000 of the simulated demands for $System_3$ and $System_4$, respectively. The point of this parameterisation is to study how the intervals between recalibration affect the prediction accuracy.

Table 5

	Prior (refer to Figure 4)	Observation (Table 3)	Number of campaigns	Demands per campaign	Total number of demands
Data1	$Prior_1$	$System_1$	50	100	5,000
Data2		$System_2$	50	100	5,000
Data3	$Prior_2$	$System_1$	50	100	5,000
Data4		$System_2$	50	100	5,000
Data5	$Prior_3$	$System_3$	25	200	5,000
Data6	$Prior_4$	$System_4$	200	200	40,000

Bayesian inference is applied to each of Data1-Data6 using the three system models – black-box, coarse, and recalibrated – starting with *consistent priors*, calculated as follows: from the prior defined for the FT-component, the marginal distribution, $f_{p_{AB}}(\bullet)$, is derived and used as marginal distribution of the UG component in the coarse model. This distribution together with $f_{p_{ROS}}(\bullet)$, defined as a Beta distribution, $B(\bullet, \alpha_{ROS}, \beta_{ROS})$, and the conditional distributions $f_{p_{ROS \wedge UG} | p_{ROS}, p_{UG}}(\bullet \mid P_{ROS}, P_{UG})$ defined as Beta distributions, $B(\bullet, \alpha_{ROS \wedge UG}, \beta_{ROS \wedge UG})$ define the joint distribution of the coarse model from which, in turn, the marginal distribution of the system *pfd* is derived and used as a prior for the black-box model.

The u-values and the PLR values were calculated for the number of system failures observed in each campaign. The values thus computed were then used to plot the respective graphs and compare the accuracy of the models on the simulated data. At the end of each testing campaign the coarse model was recalibrated using equation (9). The posteriors derived with the respective models at the end of each campaign (marginal in the case of black-box or joint in the case of coarse and recalibrated models) were then used as a prior in the next campaign, etc.

Analysing the u-plots one would notice that the u-values (on the Y-axis that is, e.g. Figure 6) do not start from 0. The u-values are constructed for the random variable 'the number of the system failures within a campaign of a given size'. Given the size of the campaign is between 25 and 200 tests for the examples shown here and the distributions of the system *pfd* are limited by the priors to small values (an upper bound of 0.02 and distribution mass often concentrated within a range much shorter than [0,0.02]) the probability of observing no system failures, P(0), may be significant, 0.75+ in the examples studied. Thus the u-plot values will be bound to be within the range [P(0), 1]. We plotted the u-plots using as an estimate of the P(0) the smallest u-value computed with each of the models.

Clearly increasing the size of the campaigns will lead to a decrease of P(0). As shown in the Appendix for a campaign size of 1000 demands, P(0) can be very close to 0.

4.3.1. Data 1

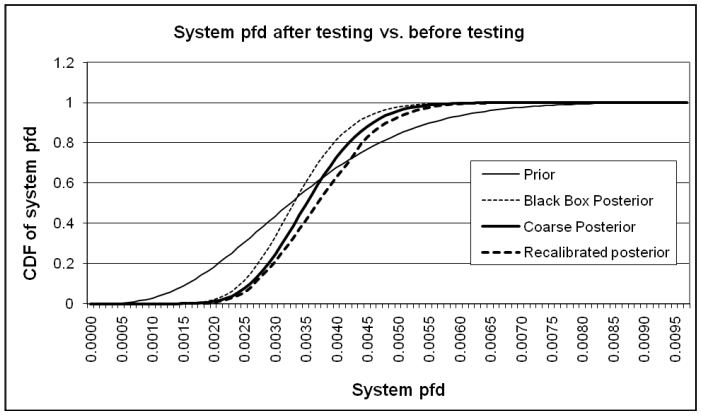

Figure 6. Distribution of system *pfd* before and after testing. There is stochastic ordering between the posterior distributions – the black-box model produces the most optimistic predictions, while the recalibrated posterior is the most pessimistic. The predictions of the coarse model are in between.

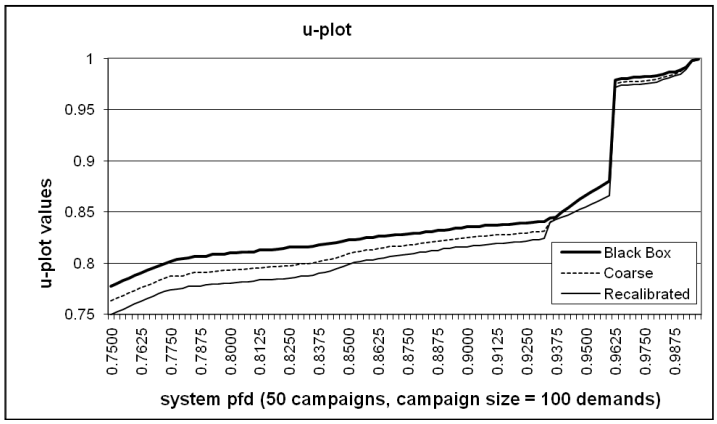

Figure 7. U-plot shows that most of the time the plots of all three models give pessimistic predictions. The ordering is the same as in Figure 6.

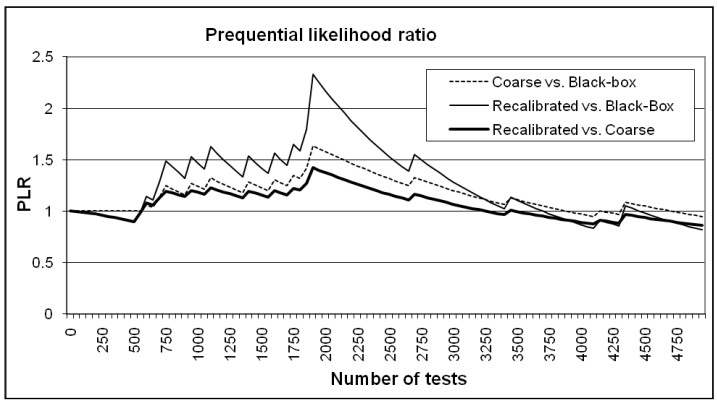

Figure 8. Prequential likelihood ratio shows minor superiority of the recalibrated model over the other two models with a tendency of all models eventually converging – they become indistinguishable, especially after 3,500 demands.

4.3.2. Data 2

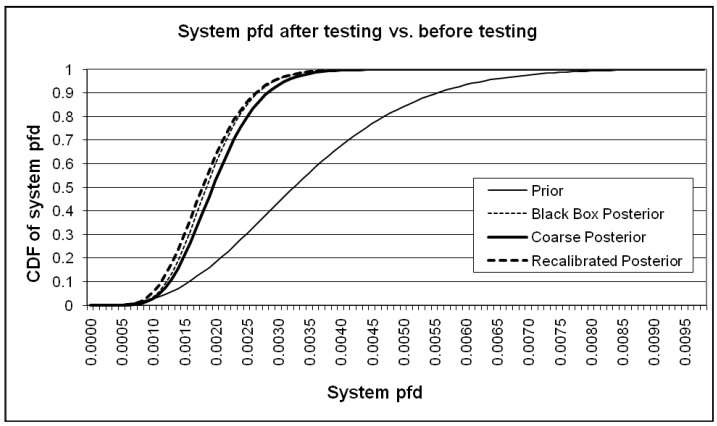

Figure 9. The plot shows that the prior chosen has been very conservative. The posteriors are very close: there is ordering between the three with the recalibrated model giving the most optimistic prediction, coarse – the most pessimistic, the black-box is in between.

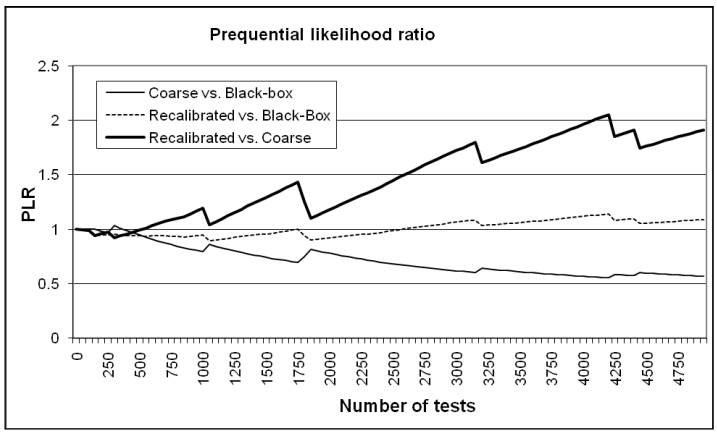

Figure 10. Interestingly, the PLR shows that the recalibrated model performs best – the difference is noticeable with the coarse model and practically negligible in comparison with the black-box model.

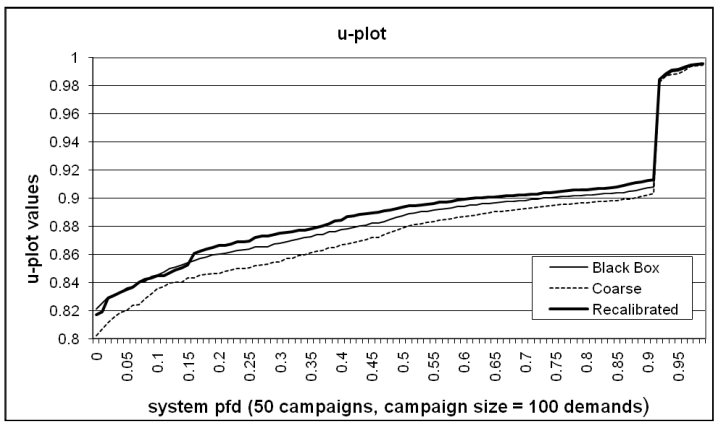

Figure 11. The u-plot reveals that all three models give generally pessimistic predictions. One may conclude, therefore, that the predictions of the recalibrated model, although most optimistic are still likely to be pessimistic and can be chosen even if one would like to be conservative about the true system *pfd*.

4.3.3. Data 3

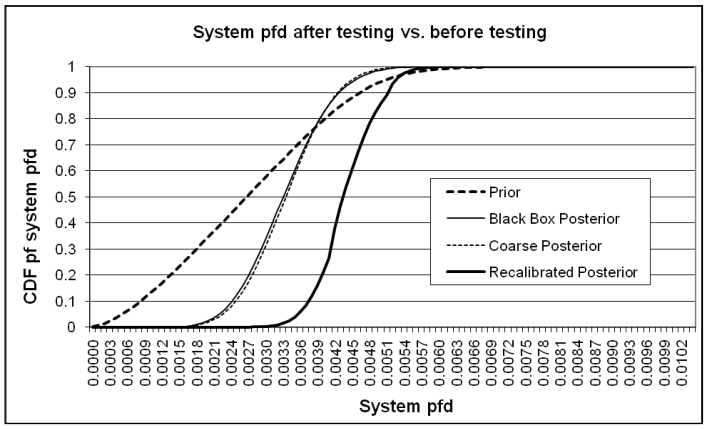

Figure 12. This plot shows a case with an optimistic prior. All posteriors are moving towards more conservative ranges. The black-box and the coarse models offer very close posteriors, while the recalibrated model is significantly different and more pessimistic.

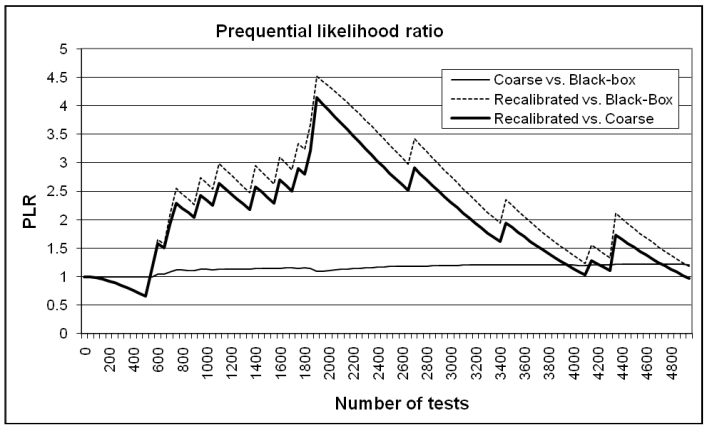

Figure 13. Now, the PLR clearly shows the superiority of the recalibrated model at the early campaigns: it outperforms both the black-box and the coarse, which are indeed practically identical (their ratio remains close to 1). The recalibrated model despite being the most pessimistic should therefore be trusted.

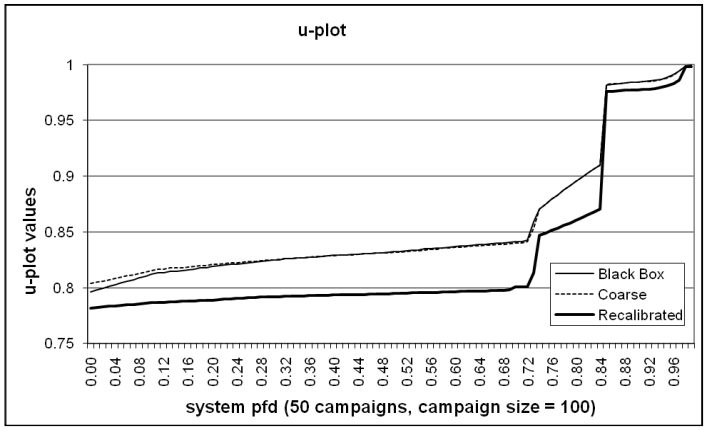

Figure 14. The observations from the PLR are now further corroborated by the u-plot. It looks like all predictions are pessimistic most of the time (the beginning of the curves which represents campaigns with no system failure – the initial curve, and a single failure – the first step of the plot), but when more than one failure occurs (the last fragments of the curves) the predictions by all three models seem optimistic. Clearly, we do not have a model that provides consistently pessimistic or optimistic predictions.

4.3.4. Data 4

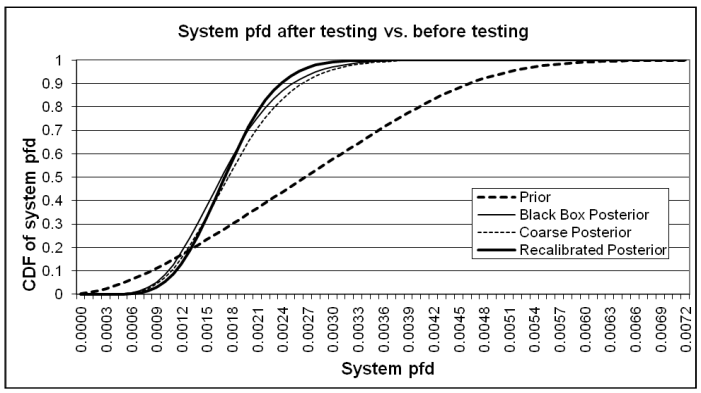

Figure 15. This case is an example of a very pessimistic prior. It is also interesting in that there is no ordering between the posteriors – the cdfs of the recalibrated posterior crosses over the posteriors obtained with the other two models. The black-box and coarse predictions seems stochastically ordered: the black-box being more pessimistic.

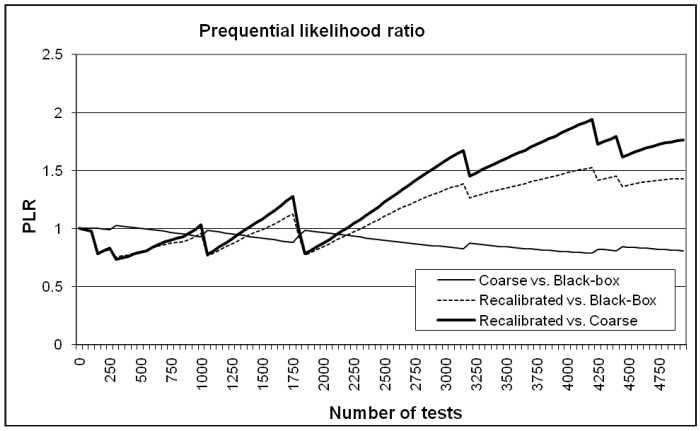

Figure 16. The PLR shows an increasing trend in favour of the recalibrated model, although the gain is minimal.

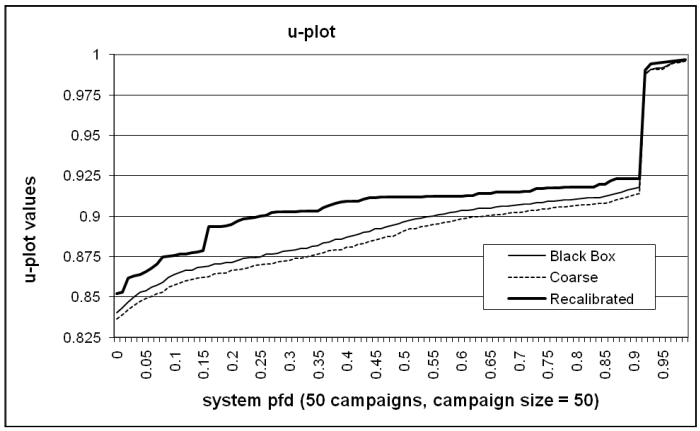

Figure 17. This plot, however, shows that the predictions of zero system failures by the recalibrated model are optimistic, noticeably so, while those by the other two models are generally pessimistic.

4.3.5. Data 5

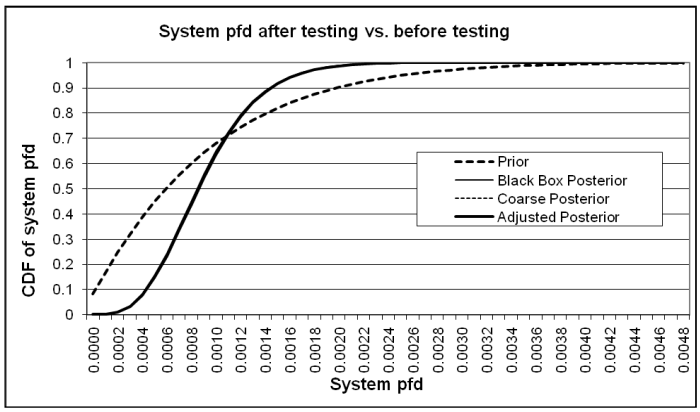

Figure 18. This case is an example of how the three models may become indistinguishable. Looking closely at the parameterisation we notice that the failures of the two channels of the FT-component (Table 4) are highly correlated. Thus, fault tolerance brings practically no benefits. Observation wise, the recalibrated model brings no benefits in comparison with the coarse model. Even the black-box model 'sees' everything that is important.

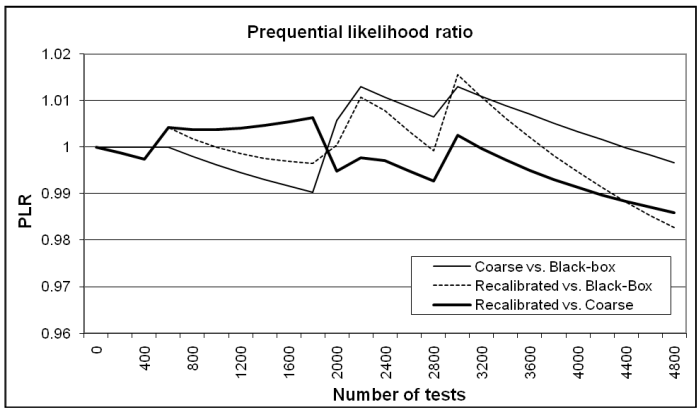

Figure 19. This plot, too, shows no difference between the three models.

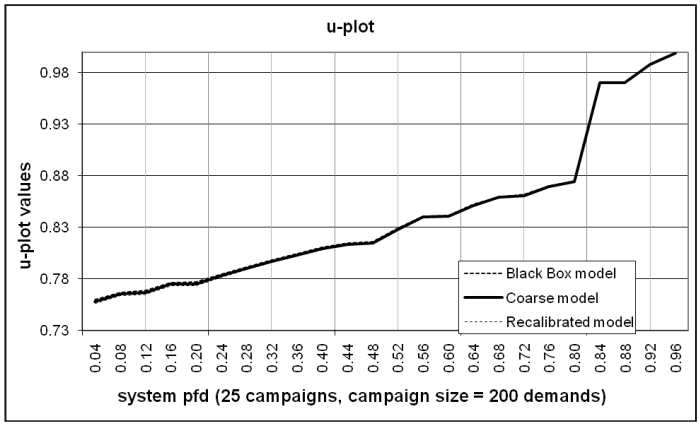

Figure 20. The same here – the u-plots of the three models are indistinguishable.

4.3.6. Data 6

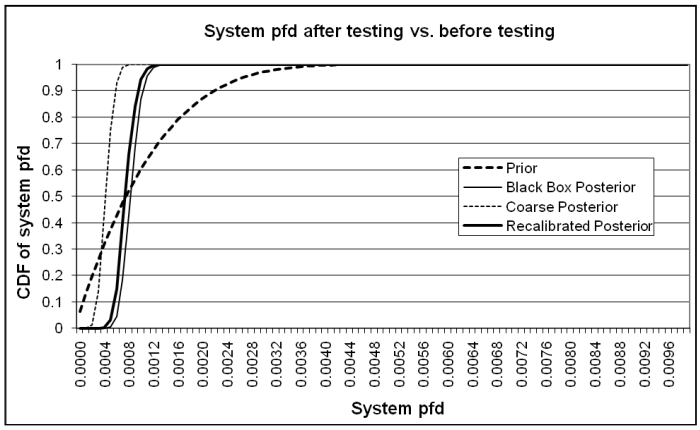

Figure 21. This example is at the other extreme. Fault tolerance works very well with the FT-component: many individual channel failures occur, and few failures of all three components. Now the recalibrated model 'sees' much more than either of the other two models, Indeed many single channel failures occur which are masked by fault-tolerance. Not surprisingly, the predictions are different. The predictions by the coarse model are much more optimistic than the other two – the black-box is more pessimistic than the recalibrated model.

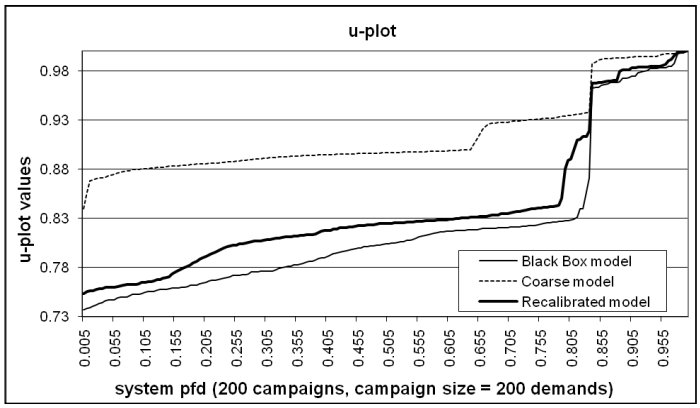

Figure 22. Clearly, the predictions by the coarse model are optimistic. For the other two models the usual pattern is observed: pessimistic u-values for the case of zero system failures.

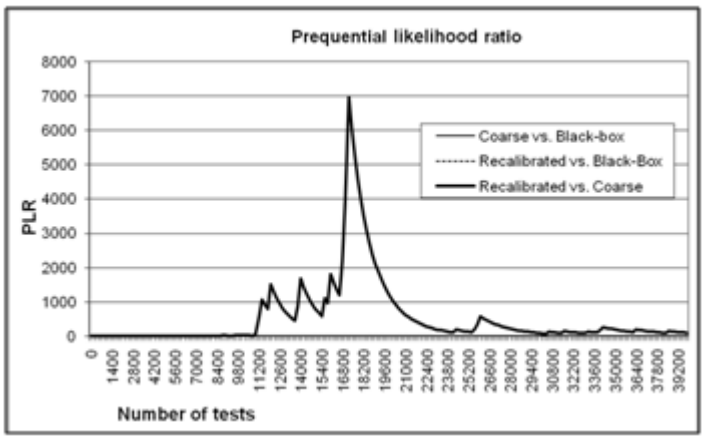

Figure 23. This plot clearly shows the drastic superiority (notice that PLR reaches values of 7000) of the recalibrated model over the coarse model.

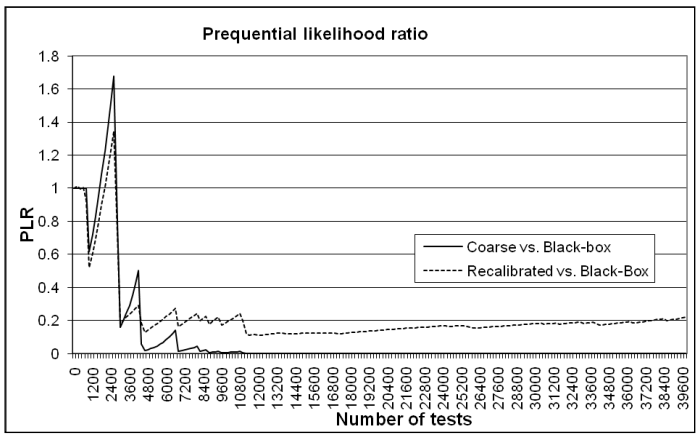

Figure 24. Here we remove the ratio between the recalibrated and the coarse model, shown above (Figure 23) and look at how the other pairs compare. Clearly the recalibrated initially performs well and outperforms the black-box model. However, as the number of tests increases the black-box becomes more accurate, although the trend suggests that the superiority may decrease.

4.4. Selecting the best prediction model

The plots shown in the previous section allow one to appreciate how the PLR and the u-plot differ between the predictions obtained with the different models. The examples also illustrate how the accuracy of the different models may change over time. The PLR and the u-plot were produced over the entire set of testing campaigns. Clearly, a more refined analysis is possible, e.g. by computing these on a sliding window of predictions as shown in Figure 25.

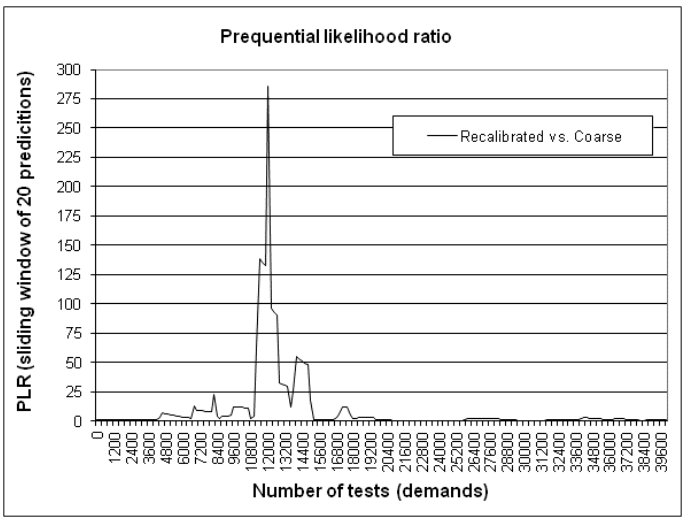

Figure 25. The PLR for Data 6 example is computed on a sliding window of 20 consecutive predictions prior to a particular point in time. The shape differs from the aggregated figure shown in Figure 23.

Clearly, Figure 25, offers a basis for one to select the best performing model. A possible selection criterion at any point in time might be the value of the PLR computed on the last 20 consecutive predictions at any point of prediction. A value of the PLR greater than 1 will give preference to the recalibrated system, while values smaller than 1 will give preference to the coarse model. One can apply different criteria to increase the confidence in the selection of the best performing model, e.g. the recalibrated can be preferred when the PLR is greater than, e.g. 10, while for lower values the evidence is insufficient for preferring the recalibrated model.

5. Discussion

The interesting question for anyone faced with using alternative models of Bayesian assessment (compared here and indeed in any other inference model) is which of the alternative predictors one should trust. In the context of safety-critical systems a prudent view would be to choose from the available alternatives the *most conservative prediction* [9]. There may be two aspects in this regard which are worth considering:

- Is the most conservative prediction conservative at all? If all available alternatives are optimistic, e.g. as a result of starting the inference with unrealistically optimistic prior, then choosing the least optimistic may still be inadequate.
- Even more interestingly – do we need to use the most conservative prediction at all? If we know that all alternatives are conservative, then any alternative, even the least conservative one, can be used for conservative predictions. An example of this possibility is given in Figure 11.

A separate line of reasoning would be to consider the *cost of being conservative*. Using a conservative prediction to demonstrate that a specified target is met is effectively equivalent to asking for the cost of the assessment to be increased (e.g. by running longer testing campaigns), which may be simply unnecessary. Had we used an *accurate* prediction (instead of the most conservative) it might be possible to demonstrate that the set target on system

reliability can be met more cheaply, e.g. with a shorter test. The numerous examples included in this paper show that the more detailed inference *in some cases* delivers more accurate predictions. The most advanced of the three models – the recalibrated – works better than the other two in most of the chosen examples. An important point to make is that the recalibrated model tends to work best when there are significant mismatches between the priors and the 'true' system reliability, especially when the channels of the FT-component are truly 'diverse', i.e. show failure diversity.

These examples seem to suggest that a feasible Bayesian inference, significantly less complex than a full inference, may result in more accurate predictions than is currently possible with simplistic models of the system such as the black box or the coarse models which ignore some details of the system structure. The proposed approach to multi-view Bayesian inference seems a significant improvement at acceptable price, compared with the Bayesian assessment typically used.

However, the results have also shown that the recalibrated model should not be trusted *a priori* and always used. The contrived examples have been chosen carefully to cover various scenarios of how the predictions obtained with different models may be related to each other. The important outcome of the work is that the known techniques for assessing the prediction accuracy of competing models seem *sufficient* for one to be able to assess objectively which of the models gives the most accurate prediction and use that model.

It is important also to appreciate the fact that *the best performing model can change over time*. This manifested itself with some of the examples although in the simulated systems their reliability remained unchanged. The reason for the change of the models' performance seems to be the mismatch between the prior and the true reliability of the assessed systems. Different models converge to true reliability with a *different speed*, which results in their prediction accuracy being different. There are cases, however, when in addition to the mismatch between the prior and the true reliability, the true reliability of the assessed system *may change over time*, e.g. as a result of a change of the usage profile. The important point here is that PLR and u-plot *seem sufficient* to identify the most accurate model whatever the reason for the different accuracies of the predictions.

As we have seen, being able to capture the trend of accuracy of the predictive models depends on the testing campaign (the intervals between applying recalibration). The *size of the campaign* should be chosen sensibly. Too long a campaign would imply that only slow changes of system reliability could be dealt with, while too short campaigns will make the techniques for assessing the accuracy (especially the u-plot) less useful. Indeed reducing the campaign size to a single demand will reduce the space of the possible observations to just 0 and 1, hence the u-plot values will be limited to a very short range, e.g. [0.9+, 1].

Another aspect that has not been discussed so far is exploring the space of possible views. Currently one of the *many possible ways* of simplifying the full inference has been chosen for the given system structure: I chose to look in a separate view at the FT-component for obvious reasons. Data 5, however, clearly shows that there is no gain from using the FT-component view. It seems pretty clear that whether the recalibration will 'work', i.e. will produce more accurate predictions than the alternative models, depends on whether it captures the 'important dependencies' between the failures of the components. With Data 6 the model of the FT-component captures observations of individual channel failure which are masked by the fault tolerance and remain 'invisible' for the other models. The model of the FT-component, however, will not see dependence (i.e. correlation – positive or negative) between the failures of ROS and one of the channels of the FT-component. Using a separate view (different from the view of the FT-component, e.g. a model of the ROS and the particular channel of the FT-component) to capture such dependence might turn out to make the predictions more accurate than can be achieved with the current choice of views. These are important aspect which will be addressed in more detail in future work. I would like to stress, however, that with many different partial views the approach presented here will also work: we will simply have to deal with a larger space of models including many different recalibrated models (which represent different set of views applied to a given full system model). The objective, however, will remain the same – compare alternative predictions of system *pfd* and choose the model (possibly the particular way of recalibration) that gives the best prediction.

Finally, in the examples used only one step of recalibration was applied – the predictions of the coarse model were recalibrated with the predicted distribution of the *pfd* of the FT component. We could, however, use multiple layers of refinement. Consider again our example in Figure 3, which was simplified by making an assumption that the component *Interface* is perfect. If such an assumption is not plausible, three different views of the system could have been defined, as shown in Figure 26.

For this model of the system we could envisage two recalibrations: The *pfd* of the FT component will be used to recalibrate the posterior of the Refined view of UG subsystem in a way similar to 3.3. The difference is that the *pfd* of this sub-system, will be derived from the joint posterior, $f_{p_{Interface}, p_{FT}, p_{Interface \wedge FT}}(\bullet,\bullet,\bullet)$ by first transforming it to a suitable form in which the system *pfd* (of a serial system), P_{UG} is explicitly used and then integrating out the nuisance parameters. The marginal distribution $f_{p_{UG}}(\bullet \mid data)$ is then used to recalibrate the coarse model as described in 3.3.

Clearly, the approach *scales up* – a structure of *arbitrary complexity* can be partitioned suitably into views which offer *chains of model refinements* by adding more detail on the system structure. Using a particular chain one can

make use of all parts of the respective model and data of the behaviour of its parts under the simplifications imposed by the particular chain.

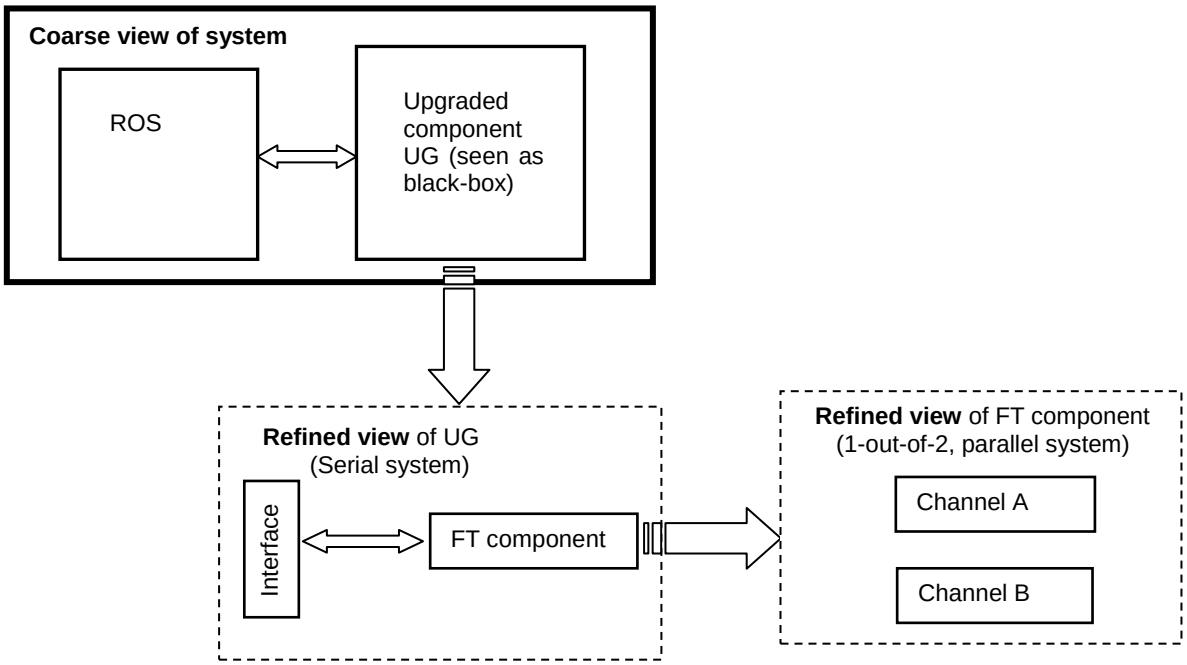

Figure 26. The example system is shown again, in which the *Interface* component can also fail. Now a new view is defined – the Refined view of UG component, which consists of two components, the Interface and the FT component modelled as a serial system, i.e. a failure of either the Interface or of the FT component will lead to a failure of UG subsystem. The third view – of the FT component – is as before; it consists of the two channels A and B, which are modelled as a 1-out-of-2 parallel system.

As an example one could "open up" the ROS, which in the presentation so far has been treated as a single component. In reality the ROS may consist of many individual components, whose behaviour might be observable, thus a complex non black-box model of the ROS can be constructed. Whether such an approach is useful will depend on the details of the specific system. As we have seen in the examples so far, predictions from the different models tend to converge over time. We assume that ROS will have been in operation for a long time, hence the black-box model of the ROS would eventually produce predictions about its *pfd* consistent with those of the more detailed models that could be used for ROS. We recognise, however, that using a more detailed model of ROS may increase significantly the space of the "*chains of model refinements*" that can be derived for the legacy system with an upgrade. The decision whether to explore this space of possibilities in search of more accurate predictions must be pragmatic. For instance, when all "*chains of model refinements*" based on ROS being treated as a black-box fail to produce an accurate prediction, then "opening up" ROS would seem justified. If with ROS as a black box, at least some of the models produce a prediction with an acceptable accuracy, opening up the ROS would seem unnecessary.

Clearly, with a complex system structure, a number of *alternative chains* might be possible. In this case, each chain will offer an alternative way of assessing system reliability and different refinements. In other words, in addition to the black-box, the coarse system model and a *single model* with recalibration as discussed in Sections 3 and 4 we will have to compare several alternative recalibration-based predictions. The task of selecting the most accurate prediction, however, does not change – simply the space of available solutions from which one has to choose gets larger.

Finally, I note that the motivation of the work presented here was based on demonstrating the difficulties in using a complete Bayesian inference with a system buit with multiple software components. In this case, the inference would rely on a model with many degrees of freedom, which is typically difficult to parameterise. As a result the benefits from a complete inference may be cancelled out by parameterisation inaccuracies. During the discussion of the paper with my colleagues we realised that the motivation could have been based on practical considerations: refining a model is only a good idea if data is available that can be made use of. From this point of view, the procedure presented here may be seen as a 'natural' way of adding details to the model when more detailed data about the behaviour of the channels of the FT component becomes available from more detailed measurements either during the testing or from operational use. The black-box view on the FT component may be useful at some point in the assessment and in the absence of information about the failures of its channels may be the only useful model of the FT component. As operational data becomes available about the behaviour of the parallel channels, however, one may decide to make use of this data (instead of throwing it away with a black-box model) and revise the model accordingly. The proposed approximate inference clearly offers one way of using the more detailed data.

6. Related research

The issue of Bayesian assessment of software built with off-the-shelf components attracted some interest. In a number of papers Kuball et al. advocated a model of software of arbitrary complex structure built with off-the-shelf software components under the assumption that the components fail independently, [10], [11]. I scrutinised their results, [12], and contrary to the assertions by those authors that the prediction errors due to the assumption of independence of failures lead to conservative predictions, observed that the sign of predictions is unknowable – the predictions that result can be overestimation or underestimation of system reliability, which poses questions about the usefulness of the models based on assuming independence of failures for practical assessment. Models of systems built with independent components have been studied very extensively and general results obtained [7].

In a previous paper, [3], we presented a Bayesian reliability assessment of 2-channel fault-tolerant software. In particular we were interested in whether a non black-box inference allows us to be confident that a predefined reliability target is met, *faster* than if this knowledge is ignored and the system is treated as a black box. We succeeded in demonstrating that using an inference procedure which takes into account the system structure and more detailed observations leads to predictions which *generally differ from the black-box predictions*. The predictions thus obtained may be more pessimistic or more optimistic than the ones obtained with the black-box inference. In general even the sign of the difference is unknowable in advance, i.e. it is impossible to know in advance what error one is making by using the black-box inference. This depends on both the prior and the observations.

Surprisingly, it turned out that some convenient priors likely to be considered by an assessor in practical assessment, because they simplify the inference, lead to counterintuitive results:

- When a Dirichlet distribution[3] is used as a prior, the black-box and white-box inferences give exactly the same posterior predictions on the probability of system failure *no matter the observations*. Hence using a white-box inference with Dirichlet gives no advantage over black-box inference, which is simpler.
- If the probabilities of failure of the two channels are assumed *known with certainty*, then the Bayesian inference may lead to counterintuitive predictions (posteriors). For instance, if no failure of either of the channels is observed (i.e. both channels passed all tests), the predicted system reliability is worse than what was assumed for it *a priori*. More precisely, the probability that the system meets a given reliability target gets lower than was assumed *a priori* after observing *no channel failure*s (i.e. the best possible outcome), which is clearly counterintuitive. The reason for this counterintuitive result is the *assumed certainty* in the reliability of the two channels and suggests that assuming certainty in the priors can have serious consequences and should be used only after a very careful justification (e.g. that it is impossible to observe no channel failure).

In an interesting study, Kristiansen et al. [13], established empirically that for systems built with software components the system dependability is typically more sensitive to dependencies between parallel components than to dependencies between serial components. They advocate that an inference procedure must account explicitly for dependencies between the parallel components, while those between the serial components can be ignored without significant loss of accuracy. The choice of the refined model of the FT-component happens to follow these authors' recommendation. Whether the empirical observation by Kristiansen et al. is universally true is unclear.

The proposed approach based on recalibration is somewhat similar to the Bayesian procedure developed in [14] for components with *independently distributed pfd*s, in which the authors use a combination of black-box inference on system reliability and inferences on components' reliability in several steps: after the posterior is derived using a black-box model (called a preliminary posterior), using the system configuration (the structure function) to derive a *consistent prior* for the components' reliability. Then the data is used to derive posteriors for the components' reliabilities, which are then used to derive the final posterior system reliability. This procedure seems to use what we called recalibration from the system model to the component models (to define priors for the components consistent with the preliminary system posterior) and back from the component models to derive the final system posterior. As [7] suggests the procedure of propagating (i.e. recalibrating) between the system/component models is quite generic and a number of variations for systems built with components with *independently distributed pfd*s have been developed since the publication of [7]. The difference of the approach proposed here in comparison with these previous models is that it does not require that the *pfd*s of the components be independently distributed.

The work presented here is also related to Bayesian Belief Nets (BBN) which have attracted interest in the past decade. The multi-view approach to assessment is a way of dealing with complexity. BBNs address this problem of complexity using *conditional independence* assumptions, which need to be justified. The inference procedure presented here does not depend on making assumptions of conditional independence. Instead of building a model of the entire system (i.e. a single BBN which involves a graph which shows the assumed conditional independencies between the variates) it deals with the multiple *partial* views in which a subset of the variates are integrated out. The approach with BBNs is different – the BBN structure states explicitly the conditional independence properties between the variates. Once the conditional independence assumptions are made they will stay in the model whatever the data. Wright and Littlewood have demonstrated that with a given set of conditional independence

[3] The Dirichlet distribution is the multivariate generalisation of the Beta distribution and forms a conjugate family with the multinomial likelihood (3). If the prior is defined as a Dirichlet distribution, the posterior is also a Dirichlet with parameters that can be easily computed from the parameters of the prior and the observed counts of failure and success of the components.

assumptions a number of different BBNs can be built, all consistent with the stated assumptions. The process of deriving the set can be automated and thus the space of possible BBNs explored in detail. In a sense, using BBNs is an alternative to the *chains of views* described in this paper. In this sense the approach and the BBN are complementary.

7. Conclusions and future research

This paper presents an approach to Bayesian reliability assessment of a complex system, which consists of many software components. The presented results suggest that the approach can solve two problems:

- Bayesian inference can be decomposed into manageable steps, each of which deals with a partial view on the system or parts thereof with simple structure. As a result, the steps do not require complex multivariate priors. Instead, they only depend on relatively simple and intuitive priors, eliciting which is practicable;
- The accuracy of Bayesian predictive models can be controlled using well known and relatively simple techniques developed in the past. With the proposed method many alternative predictive models are built and the system is assessed using all of these. Their predictions are compared and the optimal one for the specific context is chosen, e.g. one that gives the most accurate predictions or which satisfies some additional constraints, for instance giving the least pessimistic predictions among the available alternatives.

The method is illustrated on a range of contrived examples, for most of which the method performed better than the alternative simplistic models, although its superiority is not guaranteed.

Among the areas for *future research* I envisage the following:

- Deciding on the optimal set of partial models possible with a system requires further study. Using partial models inevitably throws away some of the information about the possible dependencies between the failures of the components – this is the real price of not using the full model. Deciding *a priori*, which dependencies are important may be difficult, possibly impossible. Exploring the space of partial models may be on the other hand too costly computationally which warrants further scrutiny.
- Even with relatively simple systems like the one used in this paper to illustrate the approach, the computational complexity of the assessment increases significantly. For instance in comparison with a black-box inference, the proposed method is many-fold more expensive in terms of computation resources required. For off-line assessment the speed of the inference does not pose a problem. There are, however, many potential applications, which may require fast inference on *live data*, such as critical infrastructures, cloud computing, etc. For this category of applications the speed of the inference becomes an issue and this will be addressed in future work.
- This work has not addressed how difficult it is algorithmically to choose the best model: the selection was based on processing *visually* the u-lot and PLR. Being able to automate the selection of the best model is important, e.g. in case of continuous on-line assessment, and this concern will be addressed in future work.

Acknowledgement

This work was partially supported in by the DSIPO-3 project, CI/GNSR/5014G, funded by British Energy, UK.

The author is grateful to the anonymous reviewers who made many useful comments and suggestions on improving the presentation of the work and to his colleagues Bev Littlewood and David Wright from the Centre for Software Reliability for critically reviewing earlier drafts of the paper.

References

1. Bishop, P.G., R.E. Bloomfield, and P.K.D. Froome, *Justifying the use of software of uncertain pedigree (SOUP) in safety-related applications*. HSE Books. 2001.
2. Littlewood, B. and L. Strigini, *Validation of Ultra-High Dependability for Software-based Systems.* Communications of the ACM, 1993. 36(11): p. 69-80.
3. Littlewood, B., P. Popov, and L. Strigini. *Assessment of the Reliability of Fault-Tolerant Software: a Bayesian Approach*. in *19th International Conference on Computer Safety, Reliability and Security, SAFECOMP'2000*. 2000. Rotterdam, the Netherlands: Springer.
4. Strigini, L., *Engineering judgement in reliability and safety and its limits: what can we learn from research in psychology?* 1994, SHIP Project.
5. Johnson, N.L. and S. Kotz, *Distributions in Statistics: Continuous Multivariate Distributions*. Wiley Series in Probability and Mathematical Statistics, ed. R.A. Bradley, Hunter, J. S., Kendall, D. G., Watson, G. S. Vol. 4. 1972: John Weley and Sons, INc. 333.
6. Barlow, R.E. and F. Proschan, *Mathematical Theory of Reliability*. reprint, illustrated ed. Classics in Applied Mathematics. Vol. 17 1996: Society for Industrial and Applied Mathematics. 257.
7. Martz, H.F. and R.A. Waller, *Bayesian Reliability Assessment*. Wiley series in probability and mathematical statistics. 1982: John Wiley & Sons Inc. 745.
8. Brocklehurst, S., et al., *Recalibrating software reliability models.* IEEE Transactions on Software Engineering, 1990. SE-16(4): p. 458-470.
9. Littlewood, B. and J. Rushby, *Reasoning about the Reliability of Diverse Two-Channel Systems in which One Channel is "Possibly Perfect"* IEEE Trans. on Software Engineering, accepted for publication.
10. Kuball, S., J. May, and G. Hughes. *Building a system failure rate estimator by identifying component failure rates*. in *10th International Symposium on Software Reliability Engineering* 1999. Boca Raton, FL, USA.
11. Kubal, S., May, J., Hughes, G. *Structural Software Reliability Estimation*. in *SAFECOMP '99, 18th International Conference on Computer Safety, Reliability and Security*. 1999. Toulouse, France: Springer.
12. Popov, P. *Reliability Assessment of Legacy Safety-Critical Systems Upgraded with Off-the-Shelf Components*. in *SAFECOMP'2002*. 2002. Catania, Italy: Springer.
13. Kristiansen, M., R. Winther, and B. Natvig, *On Component Dependencies in Compound Software.* International Journal of Reliability, Quality and Safety Engineering 2010. 17(5): p. 465-493.
14. Mastran, D.V. and N.D. Singpurwalla, *A Bayesian Estimation of the Reliability of Coherent Structures.* Operations Research, 1978. 26: p. 663-672.
15. Wright, D.R., *Software Reliability Prediction*, in *Center for Software Reliability*. 2001, The City University: London. p. 269.

Appendix 1

A short summary of these methods is given here in order to make this report self-contained. The text given here is based on [8]. A more detailed study of the techniques applied to discrete models, i.e. in which the parameter of interest is the number of failures in a fixed number of demands, was undertaken by Wright [15].

A.1.1. The u-plot

Let us assume that we have a series of observations, t_1, t_2, …, t_n, of a random variable T[4]. The u-plot uses the predictor $\hat{F}_i(t)$, the estimate of the distribution $F_i(t) = P(T_i \le t)$, via:

$$u_i = \hat{F}_i(t_i)$$

where t_i is the later observed realisation of the random variable T_i. Thus u_i is the value of the cumulative distribution function (*cdf*) corresponding to the observation t_i. If the sequence of predictions $\{\hat{F}_i(t_i)\}$ is good, then $\{u_i\}$ will tend to be uniformly distributed. The u-plot is the sample *cdf* of the sequence $\{u_i\}$.

In reliability growth models the random variable of interest for which the u-plot is calculated are the *inter-failure times* (e.g. the number of demands between the failures) [8], but the method does not prevent us from using any other observable random variable.

In this case if we divide the overall number of tests, N, into 'testing campaigns', n_1, n_2, …, (e.g. of fixed size, 50, 100, 250, 500, 1000 demands), then the number of observed failures (of the system or of any subsystem thereof), m_1, m_2, …, will be realisations of a random variable, *M*, with a binomial distribution, B(n_i, *p*), where n_i is the number of tests in the campaign and *p* is the system *pfd*, which in turn is a random variable distributed as defined by the respective prior distribution of the system *pfd* (prior to the particular 'testing campaign').

Let us denote that the probability density function of the distribution of the system pfd at the beginning of the i-th testing campaign as $f_{sys_i}(\bullet)$. Then the u-plot value from this testing campaign is:

$$u_i = \sum_{k=0}^{m_i} \binom{n_i}{m_i} \int_p p^{m_i} (1-p)^{n_i - m_i} f_{sys_i}(p) dp$$

Indeed this is the *cdf* of observing up to m_i failures in n_i trials (demands) when the probability of failure in a trial is a random variable with a density function $f_{sys_i}(\bullet)$.

A.1.2. The prequential likelihood

Let us assume that the predictive distribution $\hat{F}_i(t)$ has a probability density function (*pdf*):

$$\hat{f}_i(t) = \hat{F}_i'(t)$$

For predictions T_{j+1}, T_{j+2}, T_{j+n}, the prequential likelihood (PL) is defined as:

$$PL_n(j, \boldsymbol{t}) = \prod_{i=j+1}^{j+n} \hat{f}_i(t_i)$$

A comparison of two prediction systems, A and B, over a range of predictions of T_{j+1}, T_{j+2}, T_{j+n} can be made via their PL ratio:

$$PLR_n(j, \boldsymbol{t}) = \frac{\prod_{i=j+1}^{j+n} \hat{f}_i^A(t_i)}{\prod_{i=j+1}^{j+n} \hat{f}_i^B(t_i)}$$

It has been shown that if $PLR_n(j, \boldsymbol{t}) \to \infty$ as $n \to \infty$ prediction system B is discredited (i.e. has worse performance) in favour of system A. It is worth pointing out that while the u-plot indicated the type of inaccuracy – pessimistic or optimistic – the $PLR_n(j, \boldsymbol{t})$ just indicates which of the two predicting systems gives more accurate predictions without indicating the sign of the error of either of the predicting systems.

In our case we can see the testing as a series of campaigns and use the *probability of the particular observation*, i.e. the number of successes/failures, during the campaign as $\hat{f}_i(t)$, from which to calculate the $PLR_n(j, \boldsymbol{t})$.

Similarly to the u-plot, we will divide the overall testing into testing campaigns with n_1, n_2, …, etc. demands in each of them and the number of observed failures during these campaigns, m_1, m_2, …, etc. . Each of the compared models

[4] In [8] the description is given for 'next time to failure', but applies for any random variable.

– black-box, coarse, and recalibrated – models will have produced a distribution of the system *pfd*, which is denoted as $f^{j}_{sys_i}(\bullet)$.

In detail, for each of the campaigns, *i*, for model, *j*, we compute pr_i^j :

$$pr_i^j = \binom{n_i}{m_i} \int_p p^{m_i} (1-p)^{n_i - m_i} f^{j}_{sys_i}(p) dp ,$$

The new values (to be precise the ratio of the respective pr_i^j s) are counted in computing the value of the prequential likelihood accumulated up to campaign *i*.

A.1.3. Illustrations

How the two techniques work is illustrated here with two examples, called Case 1: optimistic prior, and Case 2: pessimistic prior. Details of the particular data used to compute the plots are omitted as the purpose of the plots is purely illustrative.

One can appreciate how different the two cases are. The u-plot with Case 1 is above the unit slope, which is an indication that the predictions obtained are systematically optimistic. The u-plot in Case 2 is under the unit slope which indicates that the predictions are systematically pessimistic.

The PLR plot reveals that in Case 1 we can barely prefer one prediction to another – the black-box is the best (but the difference is not really significant – the ratio computed on a relatively large number of campaigns remains modest), the coarse model is marginally better than the recalibrated. The recalibrated in this case turns out to be the worst. The fact that we can hardly differentiate between the models is clear from the u-plot, too – the u-plots are very close to each other.

A slightly different picture is revealed by Case 2. The u-plots show that the predictions of all three models are pessimistic, but the degree of pessimism is different: the coarse model is the least pessimistic, while the recalibrated is the most pessimistic. The PLR plot reveals further details. At the beginning, up to the 20th campaign the black box prediction is the best – it outperforms both the coarse and the recalibrated model. Starting from the 21st campaign, however, the coarse model outperforms the black-box model and there is a clear trend for the superiority to increase, i.e. the coarse system model best captures the system behaviour.