Assessing the predictive performance of creep models using absolute rather than squared prediction errors: an application to 2.25Cr-1Mo steel and 316H stainless steel

Mark Evans

Published online: 16 Oct 2023.

Submit your article to this journal ↗

Article views: 10

View related articles ↗

View Crossmark data ↗

Taylor & Francis
Taylor & Francis Group

Assessing the predictive performance of creep models using absolute rather than squared prediction errors: an application to 2.25Cr-1Mo steel and 316H stainless steel

Mark Evans

Institute of Structural Materials, Swansea University, Swansea, UK

ABSTRACT
A reliable means of assessing the accuracy of a creep model's predictions is fundamental to safe power plant operation. This paper introduces a method of decomposing the mean absolute prediction error for such a purpose to overcome the limitations that are inherent in the traditional approach of squaring prediction errors to prevent over and underestimates of life offsetting each other. When this method is applied to 2.25Cr-1Mo steel and 316 H stainless steel, it was found that squared errors leads to overestimates of the average prediction error associated with a particular creep model, and it also dramatically underestimates the proportion of this error that is systematic in nature. These differences were more noticeable for 316 H stainless steel.

Introduction

It is important to be able to predict the creep life and other creep properties of materials used in power plants and aero engines. When this can be done with a high degree of confidence, the results can potentially be used to justify the continued use of ageing power plants beyond their original design lives – as a short-term solution to potential energy gaps, for example – or increase operating temperatures to raise efficiency levels. 2.25Cr-1Mo is a main stay steel used for structural components operating at high temperature within such ageing power plants – where the usual service conditions for heater tubes are around 840 K and 35 MPa. 316H stainless steel is used when higher operating temperatures are required. Yet it has proved very challenging to predict the service life of these materials at such conditions using just the results from accelerated tests (tests done at much higher stresses and temperatures).

Consequently, one of the aims of the European Creep Collaborative Committee (ECCC), established in 1992, was to develop techniques for assessing creep data in Europe. In particular, the Technical Working Group (WG1) of the ECCC developed methods and guidelines for the assessment of creep data [1b – e]. The first creep data assessment exercise undertaken by this working group examined a single-cast of 2.25Cr-1Mo steel [1,2]. This exercise involved the application of 16 creep models by nine analysts and this resulted in the development of the Z parameter (among other statistics) to provide a measure of how well a creep model performs in predicting creep properties such as strain, time to failure and time to various strains. To construct this Z parameter, Holdsworth et al. [2] first defined the residual log time as

$$e_i = A_i - P_i \qquad (1a)$$

where $A_i = \ln(t_F)_i$ is the natural log of the time at which the ith test specimen fails (of which there are n such failure times in a data set) and P_i is a creep model's prediction of A_i. Note that in what follows A_i could equally stand for any other creep property such as the minimum creep rate, time to various strains or even strain itself. The reason for working with natural logs is that this residual log time is then approximately equal to the percentage difference between the actual time to failure and a model's prediction of it. This approximation is very good for percentages of around 30% or less (the smaller the better). Consequently, in this paper e_i will be referred to as the percentage prediction error. This percentage scaling enables comparisons to be made of different creep properties which have different units of measurement. Next, the authors compute the standard deviation in residual log time (labelled $S_{A\text{-RLT}}$ by Holdsworth et. al.)

$$S_{A-RLT} = s_e = \sqrt{\frac{\sum_{i=1}^{n} [e_i - \bar{e}]^2}{n}} \qquad (1b)$$

where $\bar{e}$ is the mean percentage error in predicting the log failure time associated with all tested specimens.

Note the denominator in Eq. (1b) is n, rather than the more usual *n*-1, and so if applied to a small sample of data would produce a biased estimate of the true or population standard deviation. The Z parameter is then defined as

$$Z = e^{2.58 s_e} \qquad (1c)$$

Ideally, for single-cast assessments, Z should be less than or equal to 2, whereas $Z \geq 4$ is unacceptable [1,2]. This was determined by looking at the Z values associated with the best performing creep models. For multi-cast assessments, even the best performing creep models studied by Holdsworth et al. [2] could not achieve Z values below 6–7. These values can therefore be used for benchmarking purposes. If a creep model is assumed to predict the median time to failure at a given test condition, and that failure times at a given test condition follow a normal distribution, and if s_e is also independent of test conditions, then 99% of all failure times will be within the range

$$\frac{P}{Z} \rightarrow ZP \qquad (1d)$$

no matter what the test condition is. Thus, a Z value of 2 means the predictive accuracy of the creep model is such that there is only a 0.5% chance that failure times will be more than two times the model's prediction and there is only a 0.5% chance of failure times being lower than half the model's prediction. This is the recommended level of predictive accuracy according to Holdsworth under these assumptions (for single batches of material). However, this simple rule can be misleading as it is possible for a creep model to produce a Z value of 2 or less but produce predictions that are on average incorrect – i.e. produces systematic and biased percentage prediction errors, rather than random percentage prediction errors.

Therefore, an approach taken by Evans [3] was to assess creep model adequacy using the approximate mean percentage squared error (MPSE)

$$MPSE \cong \frac{1}{n} \sum_{i=1}^{n} [e_i]^2 \qquad (2)$$

The advantage of this approach is that the MPSE can be decomposed along the lines suggested by Granger and Newbold [4] into a part that is systematic and another part that is random in nature – so overcoming the above problem with using Z. However, issues remain. First, an absolute value-based measure such as the mean percentage absolute error (MPAE) is much more interpretable. Whilst taking the square root of the MPSE (to give the RMPSE) helps with interpretation by converting the MPSE into the same units as e (i.e. to an average percentage error), this can still give very misleading assessments of a creep models adequacy. This is because the percentage prediction errors are squared and this makes the MPSE, and thus the RMPSE, very sensitive to the presence of a few outliers or poor predictions. Thus, there is the potential for the RMPSE to underestimate the predictive performance of a given creep model. This is not true of absolute percentage errors which are more robust to such outliers.

The aim of this paper is therefore to assess whether the use of the MPAE (and its decomposition into random and systematic error components) along with Z gives a better and more meaningful measure of creep model adequacy compared to using RMPSE and Z in predicting failure times using data on 2.25Cr-1Mo steel and 316 H stainless steel as test beds. To this end the paper is structured as follows. The next section summarises the data sets on these two steels. There then follows a section showing how to decompose the MPSE and the MPAE. In the results section these statistics are calculated using a parametric creep model to illustrate the misleading conclusions that can be drawn from the RMPSE. The paper concludes with recommendations for future work.

The data

This paper makes use of the information in Creep Data Sheets 3B, 50A and 14B, published by the Japanese National Institute for Materials Science (NIMS) [5–7]. These have extensive data on 12 batches of 2.25Cr-1Mo (according to ASTM A 387, Grade 22) steel where each batch has a different chemical composition that underwent one of four different heat treatments – details of which are given in [5]. This paper makes use of just one of these batches, the MAF batch, which was in tube form that had an outside diameter of 50.8 mm, a wall thickness of 8 mm and a length of 5000 mm with a chemical composition of: Fe – 2.46 Cr – 0.94 Mo – 0.1 C – 0.23 Si – 0.43 Mn – 0.011 P – 0.009 S – 0.008 Ni – 0.07 Cu – 0.005 Al. Specimens for creep testing were taken longitudinally from this material. Each test specimen had a diameter of 6 mm with a gauge length of 30 mm. The creep tests were obtained over a wide range of conditions: 400 MPa – 22 MPa and 723 K – 923 K. For the MAF batch (and only this batch) both minimum creep rates and time to failure measurements were recorded, together with the times to attain various strains −0.005, 0.01, 0.02 and 0.05. Figure 1 plots the creep failure times obtained for this MAF batch at the different stresses and temperatures used. The relationship between time to failure and test conditions is quite complicated for this batch – which has made it very difficult to model and predict such failure times using parametric creep models. In particular, the slope of a visually drawn curve changes dramatically at
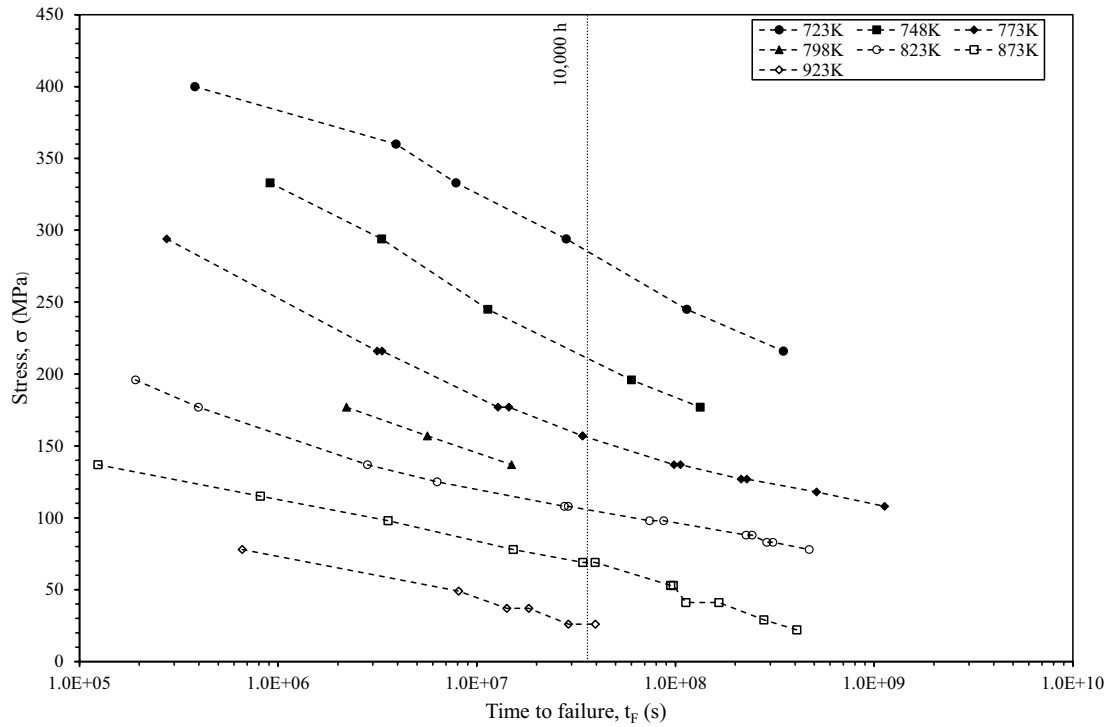
Figure 1. Relationship between stress, temperature, and time to failure for the MAF batch of 2.25Cr-1Mo steel contained within NIMS creep data sheets 3B &50A [5,6].

different test temperatures, so that some of the iso-thermal lines have more than one inflection point.

The second set of data used in this paper is on type 316 H stainless steel (18Cr-12Ni-Mo with up to 0.08% C). This data set produced by NIMS [7] reported failure

times for two batches of plate at 873 K through to 1123 K and are shown in Figure 2. These batches were labelled by NIMS as AaA to AaB. These plate specimens were hot rolled and held at 1323 K for either 40 mins or 80 mins before water quenching. More details of the
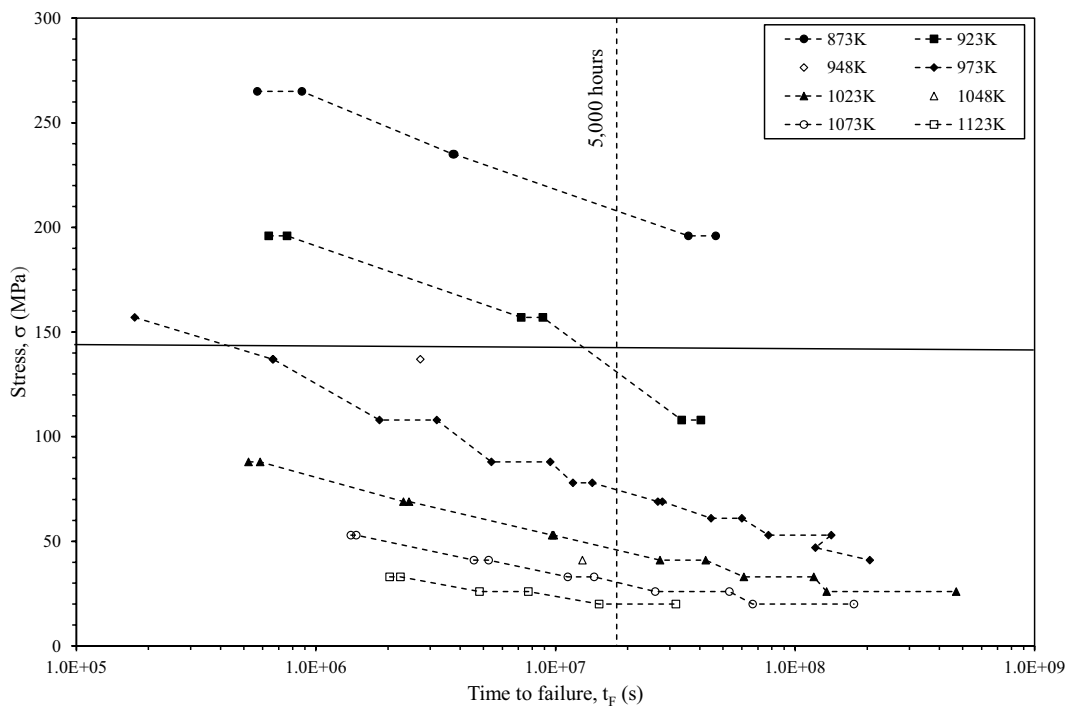

Figure 2. Relationship between stress, temperature, and time to failure for 316H stainless steel plate contained within NIMS creep data sheets 14B [7].

chemical composition and heat treatments for these two batches are contained in Reference [7].

Evaluation statistics

Limitations of the Z parameter

As mentioned in the introduction section, Z should be less than or equal to 2 for an acceptable level of predictive accuracy to have been achieved for a single batch of material. However, this simple rule can be misleading as it is possible for a creep model to produce a Z value of 2 or less but produce predictions that are on average incorrect – i.e. produce systematic and biased percentage prediction errors, rather than random percentage prediction errors. This can be seen by applying the rules of expected values through Equation (1a) to give

$$E[A] = E[P] + E[e] \qquad (3a)$$

where E stands for the expected value. An expected value is simply the population mean. There is no reason to suppose a creep model will produce values for P that yield an average or expected value for e that is zero. To see the conditions under which a creep model will produce $E[e] = 0$, i.e. will not produce systematic errors when $Z \leq 2$, rewrite Equation (1a) as

$$A_i = \alpha + \beta P_i + \varepsilon_i \qquad (3b)$$

where ε_i is a random disturbance term. If α and β are estimated using linear least squares, then

$$\hat{\beta} = \frac{s_{AP}}{s_P^2} = r\frac{s_A}{s_{AP}} \quad \text{with} \quad \hat{\alpha} = \bar{A} - \hat{\beta}\bar{P} \qquad (3c)$$

where

$$r = \frac{\sum_{i=1}^{n}[A_i - \bar{A}][P_i - \bar{P}]}{\sqrt{\sum_{i=1}^{n}[A_i - \bar{A}]^2 \sum_{i=1}^{n}[P_i - \bar{P}]^2}};$$

$$s_P^2 = \frac{\sum_{i=1}^{n}[P_i - \bar{P}]^2}{n}; s_A^2 = \frac{\sum_{i=1}^{n}[A_i - \bar{A}]^2}{n};$$

$$s_{AP} = \frac{\sum_{i=1}^{n}[A_i - \bar{A}][P_i - \bar{P}]}{n} \qquad (3d)$$

and where $\bar{P}$ is the sample average of all the log failure time predictions and $\bar{A}$ is the sample average of all the log failure times. The hat above α and β denotes that Equation (3c) gives a sample estimate of the population values for these parameters.

Then if $\hat{\alpha} = 0$ and $\hat{\beta} = 1$, $A_i - P_i = \varepsilon_i$, and so ε_i must equal e_i for Equations (1a) and (3b) to be consistent with each other. Further, by least squares construction (i.e. given Equations (3c-3d)), the average value for ε_i will be zero and so $E[\varepsilon] = 0 = E[e]$. So only when $\alpha = 0$ and $\beta = 1$ will a creep model produce predictions that are on average equal to the actual values and so have no systematic errors. Ideally then, an acceptable creep model should have $Z \leq 2$ and also $\alpha = 0$ and $\beta = 1$.

The MPSE and its decomposition

Evans [3] has decomposed the percentage prediction error e along the lines suggested by Granger and Newbold [4]

$$\begin{aligned} \text{MPSE} &\approx \frac{1}{n}\sum_{i=1}^{n}[e_i]^2 \\ &= \left(\hat{\alpha} + \left(\hat{\beta} - 1\right)\bar{P}\right)^2 + (s_P - rs_A)^2 \\ &\quad + \left(1 - r^2\right)s_A^2 \end{aligned} \qquad (4a)$$

The first bracketed term on the right-hand side in Equation (4a) is that part of the MPSE that is due to the average percentage prediction error and so is often termed the bias error. This is a systematic error present in the predictions as a positive value implies that the creep model predicts incorrectly on the average. This follows from Equations 2 that implies $\left(\hat{\alpha} + \left(\hat{\beta} - 1\right)\bar{P}\right)^2 = (\bar{A} - \bar{P})^2$. The second bracketed term on the right-hand side in Equation (4a) is due to the regression coefficient $\hat{\beta}$ differing from one, or the regression error for short. This follows from the fact that $s_P - rs_A = s_P(1 - \hat{\beta})$. This is also a type of systematic error that is present in the predictions, because $\beta < 1$ indicates that the creep model systematically underestimates the failure time below $\bar{A}$ and overestimates above $\bar{A}$ (the opposite is true when $\beta > 1$). For this reason, it is also referred to as the proportionality error. The last term on the right-hand side of Equation (4a) is due to random prediction errors because $(1 - r^2)s_A^2 = \sum_{i=1}^{n}\varepsilon_i^2/n$ and ε is pure random disturbance term. That is, the last term is just the variance in the random disturbance term (given the mean for ε is by construction zero). Dividing Equation (4a) throughout by the MPSE gives

$$\begin{aligned} 1 &= \frac{(\bar{A} - \bar{P})^2}{MPSE} + \frac{(s_P - rs_A)^2}{MPSE} + \frac{(1 - r^2)s_A^2}{MPSE} \\ &= U_M + U_R + U_D \end{aligned} \qquad (4b)$$

Consequently, $U_M + U_R$ equals the proportion of the MPSE that is systematic in nature and U_D is the proportion of the MPSE that is random in nature. An acceptable creep model therefore needs to have $Z \leq 2$, with U_D making up most of the mean squared percentage error.

Note that in this decomposition, the percentage errors are squared in the construction of the MPSE. Whilst this has the advantage of being easily modified into a test of statistical significance for any differences between competing creep model predictions using well-known distributions, it does create two major problems. The first is that an absolute-value-based measure such as the mean percentage absolute error

(MPAE) is much more interpretable. Whilst taking the square root of the MPSE (or RMPSE) helps with interpretation by converting the MPSE into the same units as e, this can still give very misleading assessments of a creep model's adequacy. This is because the percentage prediction errors are squared, and this makes the MPSE very sensitive to the presence of a few outliers or poor predictions. Thus, there is the potential for the MPSE to underestimate the predictive performance of a given creep model. This is not true of absolute percentage errors which are more robust to such outliers.

The MPAE and its decomposition

This paper therefore proposes assessing a creep model's performance using Z and a decomposition of the MPAE into systematic and random components along the lines proposed by Robeson and Cort [8]. The MPAE is formally defined as

$$MPAE \approx \frac{1}{n} \sum_{i=1}^{n} |e_i| \qquad (5)$$

Robeson and Cort have recently put forward a method for decomposing this MPAE based on the same three components as those proposed by Granger and Newbold (although they use slightly different terminology). To demonstrate this decomposition, consider the hypothetical data shown in Figure 3, where the black solid line is the regression line through the raw data (shown by the black filled circles). For these data, the MPSE = 5 and the RMPSE = 2.24. However, the MPAE = 2 and so the RMPSE tends to downplay the

accuracy of the predictions, even when there are no obvious outliers as is the case for these hypothetical data.

Unlike the decomposition of the MPSE, if a prediction has a zero error associated with it, that data point will not contribute to the MPAE, which is a strong advantage over the MPSE. Because of this, it is also possible to decompose the absolute error for each data point individually and then add them up to get the decomposition of the MPAE. This is illustrated in Figure 3 for the first data point corresponding to an actual value of −2. The absolute prediction error associated with this data point is $|e_1| = |-2 - −4| = 2$ and is given by the vertical length of the longest arrowed line. This is then decomposed into three separate parts. The first component of $|e_i|$ is the absolute bias error $|BE| = |(\bar{A} - \bar{P})| = |1.5 - 0.5| = 1$. Removing this component by adding it onto the prediction of −4 takes us to the open circled data point alongside, $P_{u1} = P_1 + BE = −4 + 1 = −3$. The dashed line is the regression line through the open circled data where the other predictions have had the bias error removed in the same way. The remaining absolute prediction error is then equal to the vertical distance from the open circled data point to the 45° line and equals a value of 1 (0.62 + 0.38). This vertical distance is made up of two components. The first component is random in nature and is made up of the vertical distance between the open circled data point and the dashed regression line, $v_1 = 0.62$. The second part is a systematic error and is made of the vertical distance between the dashed regression line and the 45° line, $f_1 = 0.38$. This vertical distance only exists because the slope of the dashed
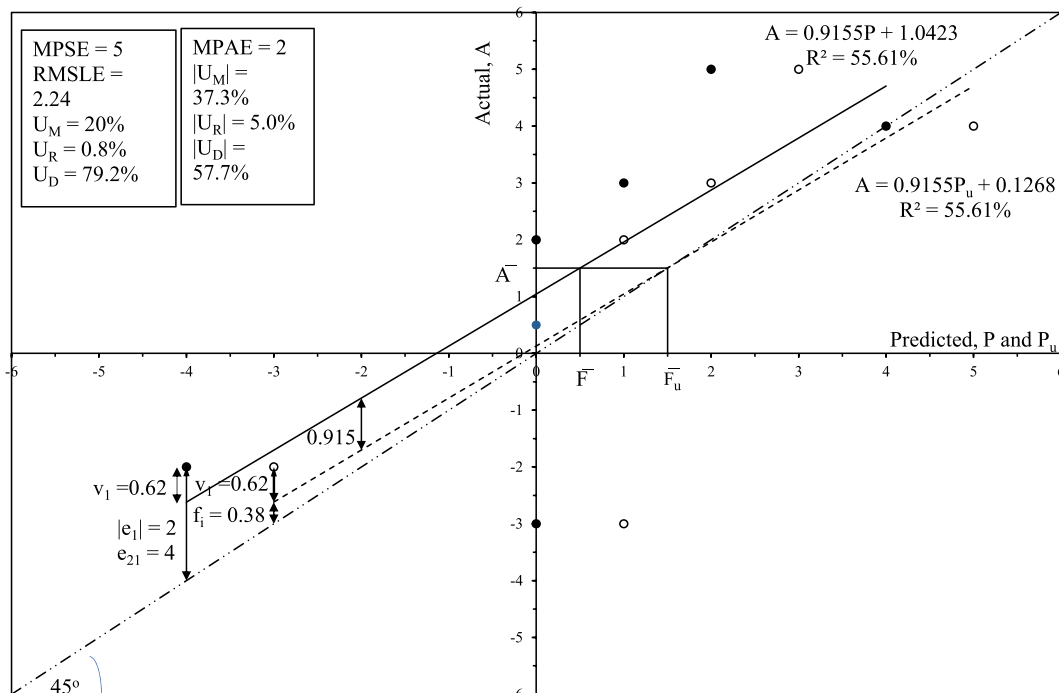

Figure 3. Illustration of the decomposition of MPAE using a hypothetical data set.

regression line is different from 1, i.e. the dashed regression line does not correspond to the 45° line. It is thus analogous to U_R above. Robeson and Willmott call this component the proportionality error. Thus, adding up the absolute value for the bias error, the random error and the proportionality error gives the absolute error for this data point, $|e_1| = |BE|+|f_1|+|v_1|$ = 1 + 0.38 + 0.62 = 2. For this data point, $|BE|$ make up 50% of $|e_i|$, the proportionality error a further 31% and the random error a further 19%.

This equivalence only works for data points above the dashed regression line and when the dashed regression line lies above the 45° line. For example, take the one but last data point corresponding to $(A_i, P_i) = [1,4]$ which does not meet this condition. Here $|e_i| = 3$, with $|f_i| = 0.127$ and $|v_i| = 2.127$ and so $|BE|+|f_i|+|v_i| = 3.25$ (which exceeds $|e_i|$). The decomposition still makes sense, however, because $f_i = -0.127$ and so $|BE|+f_i+|v_i| = |e_i| = 3$. So, the equivalence in absolute terms breaks down sometimes, but negative values cannot be allowed to offset positive values in the calculation of the average error. The solution is to divide each component by $|BE|+|f_i|+|v_i|$, and then say this ratio represents the proportion of $|e_i|$ attributable to that component. Thus, for this data point the bias equals $|e_i||BE|/|BE|+|f_i|+|v_i| = 3$ [9]$/3.25 = 0.92$. The proportionality error equals $|e_i||f_i|/|BE|+|f_i|+|v_i| = 3(0.127)/3.25 = 0.12$ and the random error, $|e_i||v_i|/|BE|+|f_i|+|v_i| = 3(2.127)/3.25 = 1.96$. These now add up to 3, i.e. to the absolute error. On this basis, the decomposition formulas are

$$MPAE_M = \frac{1}{n}\sum_{i=1}^{n}\frac{|BE|}{|BE| + |f_i| + |v_i|}|A_i - F_i| \quad (6a)$$

$$MPAE_R = \sum_{i=1}^{n}\frac{|f_i|}{|BE| + |f_i| + |v_i|}|A_i - F_i| \quad (6b)$$

$$MPAE_D = \sum_{i=1}^{n}\frac{|v_i|}{|BE| + |f_i| + |v_i|}|A_i - F_i| \quad (6c)$$

where $MPAE_M$ is the part of MPAE due to the bias – the bias error. $MPAE_R$ is the part of MPAE due to the regression line – the proportionality error. $MPAE_D$ is the part of MPAE due to the random error. The first two sum to the systematic error. One note of caution when using Equations (5,6) is that the denominator can very rarely equal zero. This can only occur when a creep model has no bias and the regression line passes through a predicted value that has no error (i.e. when $BE = 0$ and $P_i = A_i = \widehat{A}_i$, where $\widehat{A}_i$ is the value given by the regression line for A_i). This is likely to be a rare event, but in the context of creep failure time data, it is almost impossible for a model's predictions to exactly equal the recorded failure times – that are usually quoted to one decimal place. So, this issue

can be avoided by ensuring the model predictions are always expressed to more significant figures than the actual data. These equations of course give the same answer as straight addition when data points are above the dashed regression line and when the dashed regression line lies above the 45° line. So going back to the first data point

$$\frac{|BE|}{|BE| + |f_i| + |v_i|}|A_i - F_i| = \frac{1}{1 + 0.38 + 0.62}|-2 - -4|$$
$$= 1$$

$$\frac{|f_i|}{|BE| + |f_i| + |v_i|}|A_i - F_i| = \frac{0.38}{1 + 0.38 + 0.62}|-2 - -4|$$
$$= 0.38$$

$$\frac{|v_i|}{|BE| + |f_i| + |v_i|}|A_i - F_i| = \frac{0.62}{1 + 0.38 + 0.62}|-2 - -4|$$
$$= 0.62$$

which is the same as the result derived above. Like with the squared decomposition, these can be expressed as proportions of MPAE

$$|U_M| = MPAE_M/MPAE;$$
$$|U_R| = MPAE_R/MPAE; |U_D| = MPAE_D/MPAE$$
$$(7)$$

So whilst the RMPSE was only slightly inflated relative to the MPAE, when it comes to decompositions a big difference is observed using this hypothetical data set. The decomposition of the MPSE suggests the model produces prediction errors that are almost 80% random in nature. This seems at odds with the adjusted trend line (shown as the dashed line) that differs from the 45° line. This is better reflected in the decomposition of the MPAE that suggests the proportion of the model's prediction errors that is random in nature is much lower at 58%. Thus, the decomposition based on squaring prediction errors can lead to various misleading conclusions about the systematic/random nature of the model's errors in prediction. Hence, the advice is to use the MPAE, its decomposition and Z to decide which of a range of competing creep models produces the best predicted failure times.

Modelling time to failure

A selected creep model

The Orr-Sherby-Dorn (OSD) [10] parametric creep model is given by

$$\ln(t_F) = \ln(B) + n\ln(\sigma) + \frac{Q_c}{R}\left(\frac{1}{T}\right) \quad (8)$$

where T is the absolute temperature, R the universal gas constant, σ is stress and t_F time to failure. Q_c is the activation energy for creep, and B and n are further

model parameters. The model implies that there is a linear relationship between log failure time and log stress at a given temperature, and the role of temperature is then to shift this linear relationship in a parallel fashion.

When it comes to applying this model to different materials, some modifications of this equation are often required. A study of 316 H by Whittaker et al. [11] revealed changes in creep mechanism with respect to stress. They found that dislocation processes are rate controlling, with no transition to diffusional mechanisms in low stress tests. However, they found that a change in the dominant dislocation process occurs when σ falls from above to below the yield stress. Thus, when stress exceeded the yield stress, these authors stated that creep is controlled by the movement of dislocations newly generated during the plastic component of the initial strain, whereas the creep rate is determined only by grain boundary zone deformation when the stress falls below the yield stress – because no new dislocations are created during the elastic initial loading strain. It is to be expected that the activation energy for creep would be lower between grain boundaries compared to within the grains themselves. This suggested that Equation (8) requires modifying to

$$\ln(t_F) = \ln(B_o) + n_0\ln(\sigma) + \frac{Q_{c,0}}{R}\left(\frac{1}{T}\right) \quad \text{when } \sigma > \sigma_y \tag{9a}$$

$$\ln(t_F) = \ln(B_1) + n_1\ln(\sigma) + \frac{Q_{c,1}}{R}\left(\frac{1}{T}\right) \quad \text{when } \sigma < \sigma_y \tag{9b}$$

where σ_y is the yield stress, with the expectation being that $Q_{c,1} < Q_{c,0}$ and where $n_0 \neq n_1$ and $B_o \neq B_1$.

A study of the MAF batch of 2.25Cr1-1Mo by Brear [12] observed that for this material the activation energy for creep was constant with respect to test conditions, and that below some critical stress (σ^*) the resulting prolonged length of the creep tests lead to high amounts of oxidised material on the failed creep specimens that considerably weakened the creep strength of this material. This suggests that Equation (8) requires modifying to

$$\ln(t_F) = \ln(B_o) + n_0\ln(\sigma) + \frac{Q_c}{R}\left(\frac{1}{T}\right) \quad \text{when } \sigma > \sigma* \tag{10a}$$

$$\ln(t_F) = \ln(B_1) + n_1\ln(\sigma) + \frac{Q_c}{R}\left(\frac{1}{T}\right) \quad \text{when } \sigma < \sigma* \tag{10b}$$

and where again $n_0 \neq n_1$ and $B_o \neq B_1$.

Parameter estimation

As the yield stress is not published by NIMS, it must be estimated from Equations (9) and this can be achieved using a dummy variable D_1

$$\ln[t_F] = \ln[B_o] + n_o\ln[\sigma] + \lambda_1 D_1 + \lambda_2 D_1\ln[\sigma] + \frac{Q_{c,0}}{RT}$$
$$+ D_1\frac{\lambda_3}{RT} \tag{11a}$$

where D_1 equals zero when $\sigma > \sigma_y$ and unity otherwise. Thus when $\sigma > \sigma_y$, $D_1 = 0$ and Equation (11a) collapsed to Equation (9a) and when $\sigma < \sigma_y$, $D_1 = 1$ and Equation (11a) collapsed to Equation (9b) with $B_1 = B_0 + \lambda_1$, $n_1 = n_0 + \lambda_2$ and $Q_{c,1} = Q_{c,o} + \lambda_3$.

Similarly, σ^* can be estimated using

$$\ln[t_F] = \ln[B_o] + n_o\ln[\sigma] + \lambda_1 D_1 + \lambda_2 D_1\ln[\sigma] + \frac{Q_{c,0}}{RT} \tag{11b}$$

with $B_1 = B_0 + \lambda_1$, and $n_1 = n_0 + \lambda_2$

To estimate values for all the parameters in Equations (11), the following procedure can be used. First σ_y or σ^* is set equal to some starting value allowing D_1 to be quantified. Then all variables on the right-hand side of Equations (11) are fully defined so that $\ln(t_F)$ can be regressed on these right-hand side variables to obtain least squares estimates for B_0, n_0, λ_1 to λ_3 and $Q_{c,0}$. These regression equations have a residual sum of squares (RSS) associated with them – that of course is minimised by the linear least squares procedure. Then a generalised reduced gradient non-linear search technique that uses centralised numerical derivatives is used to search for values of σ_y and/or σ^* that minimise the RSS associated with the regression lines given by Equations (11). This non-linear search is carried out using Excel's Solver [13] subroutine.

Results

2.25Cr-1Mo steel

The results of estimating Equation (11b) are shown in Figure 4(a). The model's predictions are shown by the solid segmented line, and the fit to the short-term data (failure times less than 10,000 h) is good – with an R^2 value of 99.57%. The model yields an activation energy of 382 kJmol^{-1}, which is again quite close to that for lattice self-diffusion in this material. A break in the relationship seen in Figure 4(a) appears to occur at a stress of 62 MPa and the stress exponent n is much higher in the high-stress range (~-8.13), compared to ~ −2.22 in the low-stress regime. Brear [12] suggested that this difference may be due to the high amounts of oxidised material seen on the failed creep specimens at
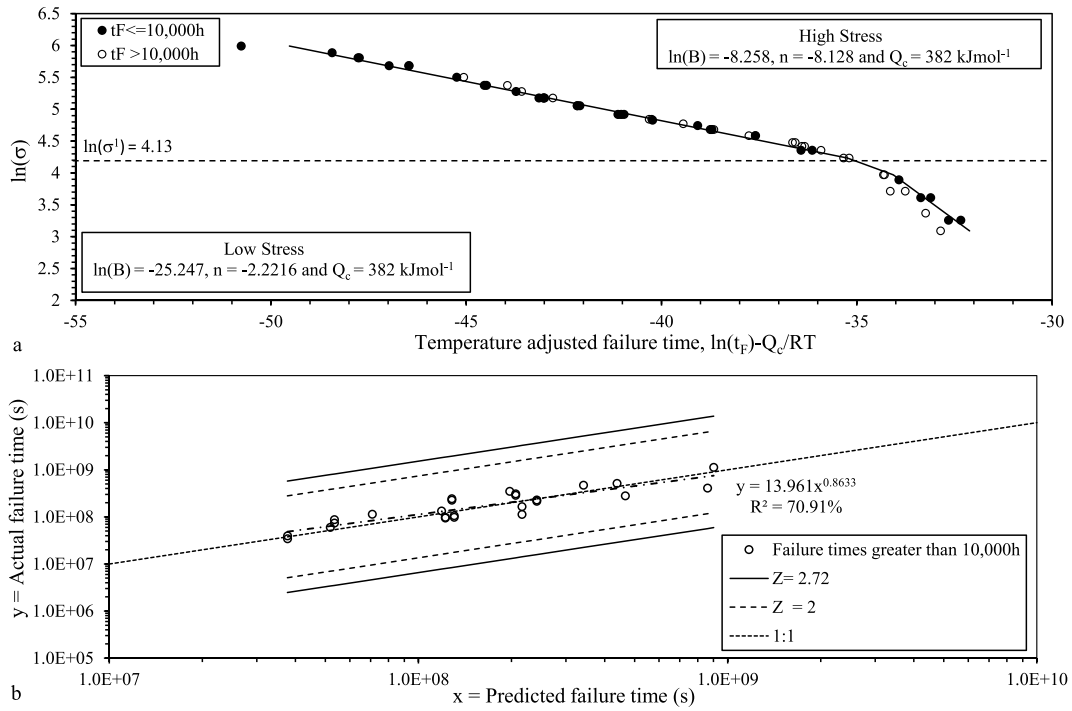
ln(σ¹) = 4.13 → $\ln(\sigma^1) = 4.13$

High Stress
$\ln(B) = -8.258$, $n = -8.128$ and $Q_c = 382$ kJmol^{-1}

Low Stress
$\ln(B) = -25.247$, $n = -2.2216$ and $Q_c = 382$ kJmol^{-1}

• tF<=10,000h
○ tF >10,000h

a

$y = 13.961x^{0.8633}$
$R^2 = 70.91\%$

○ Failure times greater than 10,000h
—— Z= 2.72
- - - Z = 2
······ 1:1

b

Figure 4. Showing (a) the OSD representation of failure times for 2.25Cr-1Mo steel, where the model is estimated using $t_F < 10,000$ h and, (b) actual v predicted t_F values beyond 10,000 h.

Table 1. Extrapolative performance of the OSD models over all temperatures for each steel.

Statistics	2.25Cr-1Mo steel	316H stainless steel
MPSE	14.81	14.68
RMPSE (%)	38.48	38.30
Z	2.72	2.74
U_M *(%)*	2.14	0.32
U_R *(%)*	9.17	29.21
U_D *(%)*	88.69	70.77
MPAE (%)	32.95	31.46
BE	0.056	0.006
$|U_M|$ *(%)*	11.65	1.43
$|U_R|$ *(%)*	17.27	43.14
$|U_D|$ *(%)*	71.08	55.43

The OSD model by Equations (9,10). MPSE defined by Equation (4a). Z defined by Equation (1c). U_M, U_R and U_D defined by Equation (4b). MPAE defined by Equation (5). $|U_M|,|U_R|$ and $|U_D|$ defined by Equation (7).

the very lowest stresses – the result of prolonged testing. Such oxidation would make the life of the material much more sensitive to a change in stress, such that a heavily oxidised specimen would not see such a big increase in life at a lower stress compared to a non-oxidised specimen.

The extrapolative performance is summarised in Figure 4(b) and in the first half of Tables 1, 2. The best fit line deviates only marginally from the ideal 1:1 line – the power term on predicted t_F is below unity (~0.86). The Z parameter equals 2.72, so that there is 99% certainty that the actual failure time at any test condition will be within 1/2.72 (= 0.37) and 2.72 times the models predicted time. So, whilst the dashed lines in Figure 4(b) are inside the solid lines, the model meets the intent of the ECCC recommendations.

The results in Figure 4(b) correspond to those in the first half of Table 1. Given the trend line has an exponent below 1 we would expect $|U_R|$ to be quite large, and Table 1 reveals that 17.27% of the MPAE is due to the best fit line in Figure 4(b) being slightly flatter than the 1:1 line. As the absolute bias error $|BE| = 0.056$ we would expect $|U_M|$ to be relatively small, and the first half of Table 1 reveals that 11.65% of the MPAE is due to the average predicted and average actual failure times being different. Thus, over all temperatures, this OSD model has a systematic prediction error of nearly 29% that is in the form of under predictions at low failure times and over predictions at higher failure times – which means the model produces slightly conservative life estimates closer to the operating conditions for this material. Consequently, most of the prediction errors made by this model are random in nature – some 71.08%. The rest of the first half of Table 1 shows the distortion in these conclusions that stem from using squared percentage prediction errors (rather than absolute percentage errors). The RMPSE of 38.48% overestimates the inaccuracy of the model's predictions as its nearly 6 percentage points above the average absolute error. More misleading is the decomposition of the MPSE. This decomposition suggested that nearly 89% of the percentage prediction errors are random in nature, which is a substantial overestimate compared to the 71% value for $|U_D|$.

The above comments relate to the model's performance over all stresses and temperatures that induce a failure time beyond 10,000 h. The first half of Table 2

Table 2. Extrapolative performance of the OSD models broken down by temperature.

	2.25Cr-1Mo steel					316H stainless steel						
	723K	748K	773K	823K	873K	873K	898K	923K	973K	1023K	1073K	1123K
MPSE	27.62	1.75	3.56	20.85	18.02	6.24	19.64	4.45	8.51	21.64	26.47	7.17
RMPSE (%)	52.55	13.22	18.88	45.66	42.44	24.97	44.32	21.10	29.18	46.52	51.45	26.78
Z	1.19	1.06	1.68	1.39	2.03	1.62	-	1.40	2.23	3.51	3.77	-
U_M (%)	99.17	98.67	6.21	93.28	63.62	100	100	100	0.74	8.92	25.03	100
U_R (%)	0.83	1.33	87.74	0.38	29.63	-	-	-	45.18	6.67	6.26	-
U_D (%)	-	-	6.05	6.34	6.75	-	-	-	54.08	84.41	68.71	-
MPAE (%)	52.33	13.13	17.37	44.10	35.00	21.20	44.32	19.00	24.95	39.22	42.16	26.78
BE	0.52	0.13	−0.05	0.44	−0.33	0.21	0.44	0.19	0.03	0.14	0.26	0.27
$|U_M|$ %	10.23	33.01	14.60	11.42	9.32	1.96	1.35	2.27	1.82	1.08	1.05	2.20
$|U_R|$ (%)	13.09	46.25	19.77	15.48	17.17	45.33	73.61	58.95	53.27	36.13	23.69	75.46
$|U_D|$ (%)	76.68	20.74	65.64	73.10	73.51	52.70	25.04	38.78	44.92	62.79	75.26	22.34

The OSD model by Equations (9,10). MPSE defined by Equation (4a). Z defined by Equation (1c). U_M, U_R and U_D defined by Equation (4b). MPAE defined by Equation (5). $|U_M|$, $|U_R|$ and $|U_D|$ defined by Equation (7).

breaks this analysis down by temperature. The MPAE is smallest at 748 K (13.13%) and largest at 723 K (52.33%). On average the model overestimates at 773 K and 873 K but on average underestimates at the other temperatures (as revealed by the sign on BE). The RMPSE is slightly pessimistic about the model's extrapolative capability at all temperatures (as shown by RMPSE > MPAE and especially so at the highest two temperatures). Apart from at 748 K, the random component of the model's absolute prediction errors is always about 65% or higher. However, at 748 K the model has a large systematic prediction error component, but thankfully the MPAE is the smallest of all the temperatures at 13.13%. Again, the decomposition of the MPSE is highly misleading. It suggests that the random component of the squared prediction errors is between 6% and 7%. But the values for $|U_D|$ are between 65% and 77% and so U_M and U_R vastly overestimate the

size of the systematic error made by this model. For the lowest two temperatures, the value for U_D cannot be quantified due to too few data points (which in turn leads to the over estimation of the systematic error using squared prediction errors).

316 H stainless steel

The results of estimating Equation (11a) are shown in Figure 5(a). The model's predictions are shown by the solid segmented line, and the fit to the short-term data (failure times less than 10,000 h) is good – with an R^2 value of 99.98%. The estimated value for the yield strength σ_Y is 142 MPa. Above this yield strength, the model yields an activation energy of 549 kJmol^{-1}, whilst below this stress the activation energy is much lower at 376 kJmol^{-1}. Whilst it is expected that the activation energy would be lower below the yield stress, the values
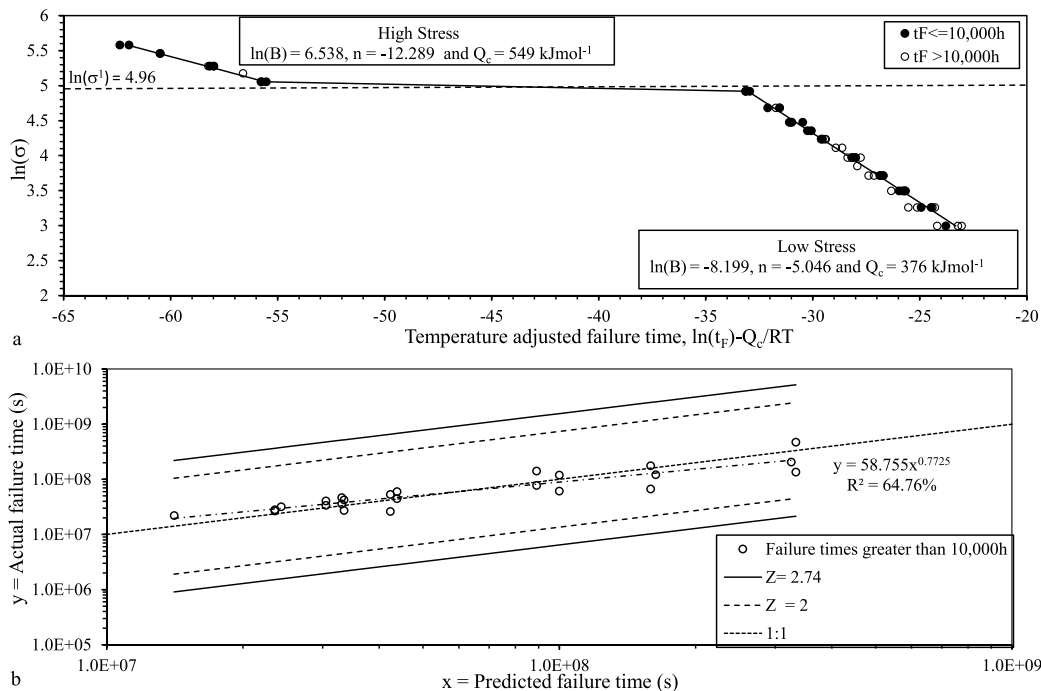

Figure 5. Showing (a) the OSD representation of failure times for 316H stainless steel, where the model is estimated using $t_F <$ 10,000 h and, (b) actual v predicted t_F values beyond 10,000 h.

are a lot higher than those quoted by Wilshire et al. [11]. – probably the result of not using the Wilshire model that normalises the stress using the high temperature tensile strengths. The stress exponent n is much higher above the estimated yield stress (~ −12.29), compared to ~ −8.29 below this stress.

The extrapolative performance is summarised in Figure 5(b) and in the second half of Tables 1, 2. The best fit line deviates from the ideal 1:1 line – the power term on predicted t_F is below unity (~0.77). The Z parameter equals 2.74, so that there is 99% certainty that the actual failure time at any test condition will be within 1/2.74 (= 0.36) and 2.74 times the model's predicted time. So, whilst the dashed lines in Figure 5(b) are inside the solid lines, the model meets the intent of the ECCC recommendations.

The results in Figure 5(b) correspond to those in the second half of Table 1. Given the trend line has an exponent below 1 we would expect $|U_R|$ to be quite large, and Table 1 reveals that 43.14% of the MPAE is due to the best fit line in Figure 5(b) being flatter than the 1:1 line. As the absolute bias error $|BE| = 0.006$ is very small, we would expect $|U_M|$ to be small as well, and the second half of Table 1 reveals that 1.43 of the MPAE is due to the average predicted and average actual failure times being different. Thus, over all temperatures, this OSD model has a systematic prediction error of nearly 45% that is in the form of under predictions at low failure times and over predictions at higher failure times – which means the model produces slightly conservative life estimates

closer to the operating conditions for this material. Consequently, just over half of the prediction errors made by this model are random in nature – some 55.43%. The rest of the second half of Table 1 shows the distortion in these conclusions that stem from using squared prediction errors (rather than absolute errors). The RMPSE of 38.3% overestimates the inaccuracy of the model's predictions as its nearly 7 percentage points above the average absolute error. More misleading is the decomposition of the MPSE. This decomposition suggested that nearly 71% of the prediction errors are random in nature, which is a substantial overestimate compared to the 55.43% value for $|U_D|$.

The above comments relate to the model's performance over all stresses and temperatures that induce a failure time beyond 10,000 h. The second half of Table 2 breaks this analysis down by temperature. The MPAE is smallest at 923 K (19.00%) and largest at 898 K (44.32%). On average the model underestimates at all temperatures (as revealed by the sign on BE). The RMPSE is slightly pessimistic about the model's extrapolative capability at all temperatures (as shown by RMPSE > MPAE and especially at the highest temperatures). Depending on the temperature, the random component of the model's prediction errors varies between 25% and 75%. Again, the decomposition of the MPSE is highly misleading. For example, at 1023 K the random component of the squared prediction errors is 84.41%, but the value for $|U_D|$ is only 62.79%. So, U_M and U_R vastly overestimate the size of the
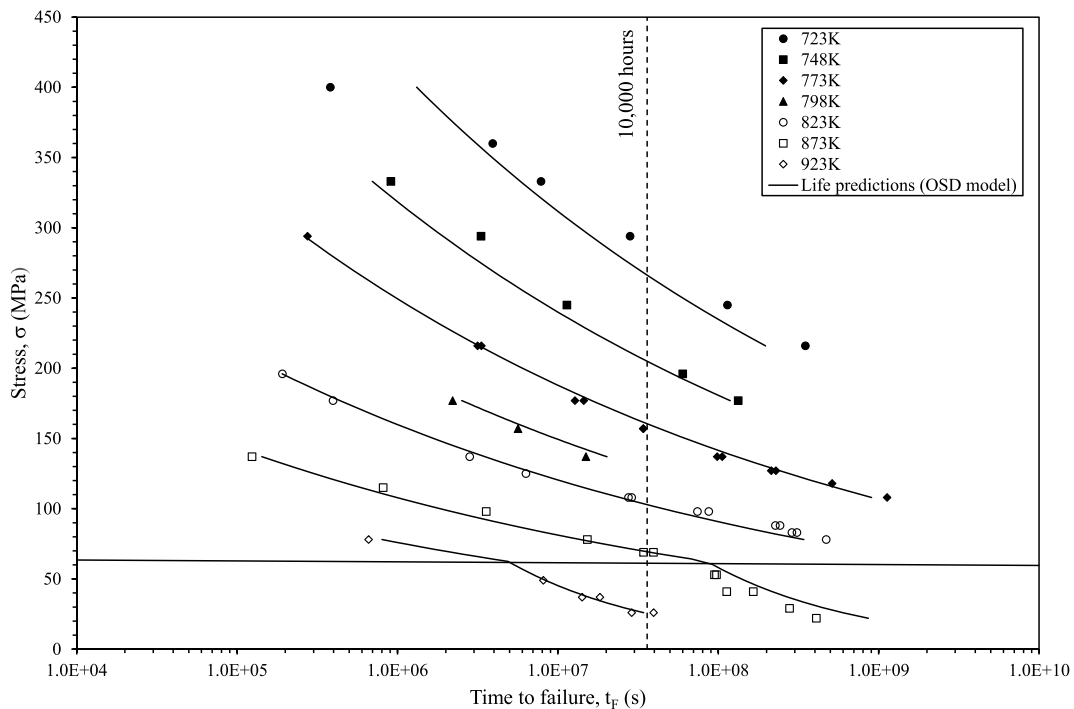

Figure 6. Showing predictions of t_F at various stresses and temperatures relative to the actual failure times (predictions based on data with $t_F < 10000h$) for 2.25Cr-1Mo steel.
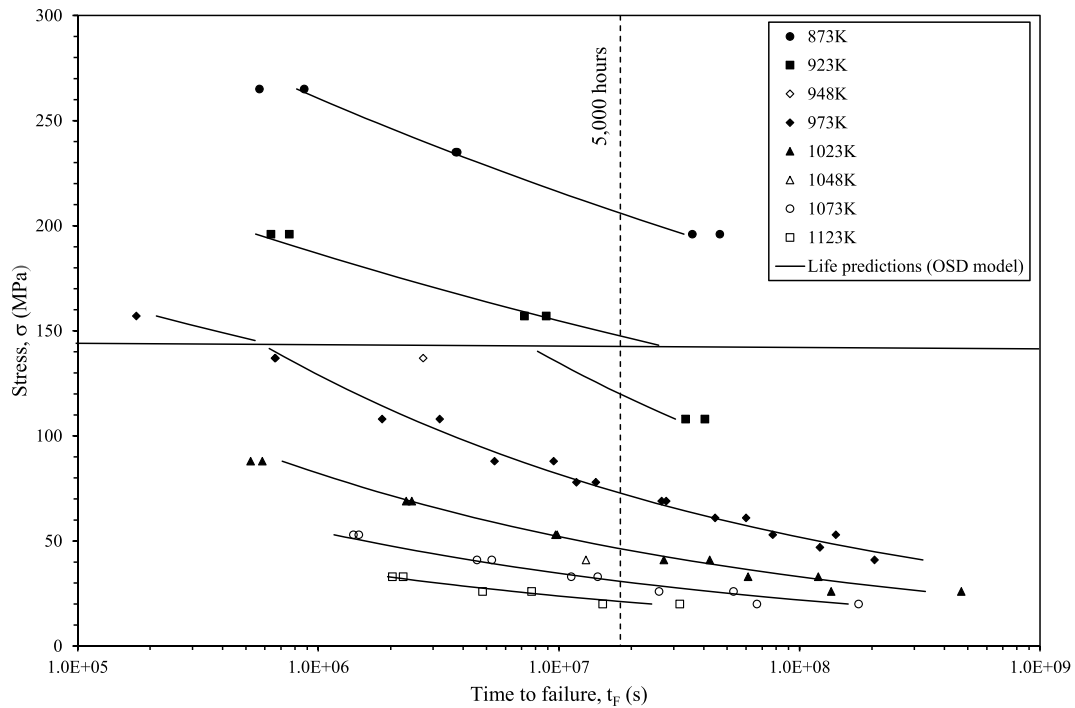
Figure 7. Showing predictions of t_F at various stresses and temperatures relative to the actual failure times (predictions based on data with $t_F < 10000h$) for 316H stainless steel.

systematic error made by this model. For the lower temperatures and the highest temperature, the values for U_R and U_D cannot be quantified due to too few data points (which in turn leads to the over estimation of the systematic error using squared prediction errors).

Material comparison

Whilst the OSD model produced very similar values for Z and the MPAE over all temperatures when applied to each material, the model clearly works better when applied to 2.25Cr-1Mo – because it has a much higher $|U_D|$ value. For 2.25Cr-1Mo steel, some 71.08% of the MPAE is random in nature compared to just 55.43% for 316 H steel. Irrespective of the temperature, the value for $|U_M|$ is between 9% and 15% for 2.25Cr-1Mo steel (except at 748 K where it is 33%). But for 316 H, $|U_M|$ never exceed 3% no matter what the temperature is, and so for this material the difference between the average failure time and the average prediction is much smaller. But for 316 H, a much bigger proportion of the MPAE is due to $|U_R|$ and so for this material there is a stronger tendency to under-predict with increasing failure times. These trends are seen in Figures 6 and 7 that show the predictions made by the OSD model in the more familiar stress-time space.

Conclusions

This article has proposed that the assessment of a creep model's predictive capability should be based on absolute rather than squared errors to avoid distortions arising from the squaring of prediction errors associated with outlying data points. It also presented a way of decomposing the average absolute error into random and systematic components to further quantify a model's performance – a procedure not used so far in creep model evaluation. When applied to 2.25Cr-1Mo and 316 H failure time data, it was found that the MPSE tended to underestimate the predictive capability of the OSD creep model. It also tended to dramatically overestimate the component of the prediction error that was random in nature in these two materials. Areas for future work include the application of these assessment statistics to other materials – especially data sets that have more heat-treated batches (so as to study the role of batch-to-batch variation) and to develop statistical tests to see whether one creep model has a better MPAE compared to another.

Disclosure statement

No potential conflict of interest was reported by the author.

References

[1] Holdsworth SR, Merckling G, ECCC developments in the assessment of creep-rupture data. In: Holdsworth SR, Orr J, Granacher J, et al., editors. Proceedings of sixth international Charles Parsons Conference on engineering issues in turbine machinery, power

plant and renewables; Trinity College, 16–18 September,; Dublin; 2003.

[2] Holdsworth SR. Developments in the assessment of creep strain and ductility data. Mater High Temp. 2004;21(1):125–132. doi: 10.1179/mht.2004.004

[3] Evans M. Estimating threshold stresses using parametric equations for creep: application to low-alloy steels. Mater Sci Technol. 2023;1–16. doi: 10.1080/02670836.2023.2198395

[4] Granger C, Newbold P. Some comments on the evaluation of economic forecasts. Appl Econ. 1973;5 (1):35–47. doi: 10.1080/00036847300000003

[5] NIMS Creep Data Sheet No. 3B. Data sheets on the elevated-temperature properties of 2.25Cr-1Mo steel for Boiler and heat exchanger seamless tubes (STBA 24). Tokyo, Japan: National Research Institute for Metals; 1986.

[6] NIMS Creep Data Sheet No.50A. Long-term creep Rupture data obtained after publishing the final edition of the creep data sheets. Tokyo, Japan: National Research Institute for Metals; 2015.

[7] NIMS Creep Data Sheet No.14B. Data sheets on the Elevated-temperature properties of 18Cr-12Ni-mo stainless steel Plates for reactor vessels (316HP). Tokyo, Japan: National Research Institute for Metals; 1988.

[8] Robeson SM, Cort JW. Decomposition of the mean absolute error (MAE) into systematic and unsystematic components. PLoS One. 2023;18(2): e0279774. doi: 10.1371/journal.pone.0279774

[9] ECCC Recommendations. Creep data validation and assessment procedures. In: Holdsworth SR, eds., ECCC. publ., (a) vol. 1: overview, (b) vol. 2: terms and terminology, (c) vol. 3: data acceptability criteria, data generation, (d) vol. 4: data exchange and collation, (e) vol. 5: data assessment, (f) vol. 6: characterisation of microstructure and physical damage for remaining life assessment, (g) vol. 7: data assessment—creep crack initiation, (h) vol. 8: data assessment—multi-axial, (i) vol. 9: component assessment.

[10] Dorn JE, Shepherd LA, What we need to know about creep. In Proceedings of the STP 165 Symposium on The Effect of Cyclical Heating and Stressing on Metals at Elevated Temperatures, Chicago, IL, USA, 17 June 1954.

[11] Whittaker MT, Evans M, Wilshire B B. Long-term creep data prediction for type 316H stainless steel. Mater Sci Eng A. 2016;552:145–150. doi: 10.1016/j.msea.2012.05.023

[12] Brear JM. A perspective on the Wilshire creep equations. Strength, Fract Complex. 2022;15 (1):79–98. doi: 10.3233/SFC-228006

[13] Microsoft Corporation. Microsoft Excel [Internet]. 2018. Available from: https://office.microsoft.com/excel.