npj | Clean Water

www.nature.com/npjcleanwater

ARTICLE OPEN

Using generalized additive models to investigate the environmental effects on pipe failure in clean water networks

Neal Andrew Barton[1], Timothy Stephen Farewell[2]✉ and Stephen Henry Hallett[1]

Predicting pipe failures using statistical modelling benefits from detailed knowledge of the conditions and circumstances which influence such failures. Incorporating this knowledge into model building improves failure predictions. In this study, we model weather, soil and hydrogeological variables in a generalized additive model for five common pipe materials separately, using partial dependence plots to understand the partial effects of each variable on pipe failure. We show how severe temperatures are associated with high pipe failure. Cold temperatures and air frost and their interaction with soils represent the key factors for pipe failures during the winter for metal pipes. Warm temperatures, high soil moisture deficit and soil movement results in higher pipe failures in asbestos cement pipes during the summer. Warm temperatures, ground movement and soil wash out, and water demand are key factors for polyvinyl chloride pipe failure during the summer. Frost is a key factor influencing polyethylene pipes during winter. An understanding of the physical principals concerning pipe failures can enable the development of more accurate models, guiding network management plans to help reduce asset leakage through appropriate interventions.

npj Clean Water (2020)3:31 ; https://doi.org/10.1038/s41545-020-0077-3

INTRODUCTION

In the UK, three billion litres of drinking water is lost daily through pipe failures[1]. Pipe failures can amass considerable financial costs from wasted water processing, proximal property and infrastructure damage, service interruption for consumers and network repairs. A comprehensive overview of the impacts of water mains failure has been previously undertaken[2]. The implications of pipe failure are significant, and with an estimated 35% future increase in UK water demand by 2050, in addition to water stress from environmental risks under future climates[3], water companies are being challenged by the industry regulator to reduce water lost through pipe failures.

Clean water assets must be managed appropriately to reduce the impacts of pipe failure. One way to manage pipe failure is through reactive maintenance, performing repairs only once they have been identified (typically when water reaches the surface). Reactive maintenance is preferably performed immediately after or within a very short time of the failure, depending on its nature (loss of water by volume is sometimes used to determine the severity of the pipe failure and how quickly the repair must be completed). Sometimes, when pipe failures are small they can remain undetected for months or even years, meaning a considerable volume of water can be lost over time[4]. To develop a cost-effective means of managing and maintaining networks, statistical models can be developed which are derived from the relationship between pipe failures and causal factors[5]. Statistical models can be used to identify pipes potentially at risk of failure, supporting operational management decisions and classifying pipes for replacement. In the last two decades, UK water companies have increased the quantity and quality of network data, enabling a wider range of data correlated with pipe failure to be considered when building statistical models. An understanding of how each variable influences pipe failure is important and recent studies have provided useful insights, exploring environmental effects on pipes[6–10].

Pipe failure mechanisms are unique for different pipe materials. The literature has reported a correlation between seasonal variation and pipe failures, as a consequence of changing weather (temperature, frost and rainfall deficit (RD)) and soil conditions (pH, ground hazards such as shrink swell, texture, moisture content)[6–8,11,12]. Failures during the winter are typically found in iron and to a lesser extent steel and ductile iron (SDI) pipes, and are associated with cold temperatures (typically below 3 °C), internal water temperature, rapid temperature transit and prolonged periods of frost[8,9,13]. During the summer a higher number of pipe failures are typically found in asbestos cement (AC) and polyvinyl chloride (PVC) pipes[6,7] and are associated with temperature and high RD which results in ground movement from soil shrink swell, associated with clay soils. Although not the same, soil moisture deficit (SMD) can be used as a corollary for RD since both are measures used to understand soil moisture content and the subsequent effects on volumetric expansion and contraction in clay soils[14]. Polyethylene (PE) is affected by environmental factors, but to a lesser extent than with other materials[11]. These studies provide useful insights on how environmental factors can influence pipe failures, however, many studies focus on only a few variables to explain pipe failures. Due to their interrelated nature, it is important to look at a wide range of soil and weather variables together to explain environmental effects on pipe failures. Furthermore, understanding the effects of the variables simultaneously, considering the regional environmental and network conditions, is important[15,16]. This understanding can be incorporated into a statistical models to improve pipe failure predictions and inform decision making[17].

This study seeks to contribute to the wider understanding of the environmental impacts on pipe failures in common pipe materials, including iron, SDI, AC, PVC (collectively unplasticised, post chlorinated and molecular orientated PVC) and PE (medium and high density). This study performs multivariate analysis using generalized additive models (GAM) to understand how the

[1]School of Water, Energy and Environment, Cranfield University, Bedfordshire MK43 0AL, UK. [2]MapleSky Ltd, 20-22 Wenlock Road, London N1 7GU, England. ✉email: tim@timfarewell.co.uk

covariates fit the model and affect pipe failure. GAMs are an approach used extensively in environmental modelling and provide great scope to model complex relationships between covariates. We look for variable dependence and interpret the effects of the covariates using partial dependence plots (PDPs). Pragmatically, visual aids can help to determine which variables have the strongest effect, are useful for interactive model building, and are easy to interpret[18]. We use data collected over 14 years for an entire network (~40,000 km of pipes) provided by a large UK regional water network.

RESULTS

The results for the covariate dependence test are shown in Table 1. A weak variable dependency is noted between temperature and days air frost, however, no strong variable dependence (>0.5) was observed. All models converged successfully suggesting the cubic penalised regression approach has generalized the model enough to reduce the effects of concurvity, if present.

PDPs show how the mean number of observed pipe failures (*y*-axis) changes over the variable interval distribution (*x*-axis). The mean centred number of pipe failures is represented by the red line, whilst the blue dotted lines indicate the 95% confidence interval. The ticks on the bottom *x*-axis show the number of observations by the variable interval. The letter on the *y*-axis represents either a smoothed variable (*s*) or categorical variable (*c*), whilst the number reported in parenthesis (e.g. Temp, 8.18) represents the effective degrees of freedom of the smoothed curves and degrees of freedom for the categorical data. The variation in mean predicted pipe failures reveals how influential the variable is on pipe failures. Therefore, we observe variables with the highest increase or decrease in mean prediction as being those which can explain pipe failures. The effects of each variable are discussed further by each material type.

Iron

The GAM PDPs for iron pipes are presented in Fig. 1 and show that all variables included in the model are significant. The highest number of pipe failures can be seen in pipes with a diameter <166 mm, pipe failures then decrease as the pipe diameter increases. The pipes age does not appear to show a distinct relationship with pipe failure, and has high uncertainty in these data suggesting either a small or no effect. Variables leading to the highest change in iron pipe failure includes temperature, days air frost, SMD and diameter band. These variables show a distinct increase in pipe failures over all or a majority of the variable interval. Iron pipe failures increase as the temperature decreases below 10 °C and with more days of air frost in a month up to ~15 days, thereafter reducing as a consequence of fewer observations and a wider confidence interval. Iron pipe failures also increase as SMD increases and is consistent with temperature where pipe failures increase above 15 °C. Soil shrink swell and subsoil type shows a small effect on iron pipe failures but suggested peat and clay soils and soil with a shrink swell potential of 12–15% volumetric expansion are associated with more pipe failures. Soil pH shows no obvious effect on iron pipe failures which is uncharacteristic since highly acidic or alkaline soils are typically associated with corrosion in metal pipes. However, pH is only one aspect of soil conditions that cause corrosion.

Asbestos cement

Figure 2 shows the GAM PDPs for AC pipes and indicates higher failure rates for pipes with a diameter < 166 mm and pipes installed between 1940 and 1961. Temperature, SMD and days air frost appear to have the largest effect on AC pipe failures, where more failures occur when temperatures are high (>15 °C), SMD is high (>100 mm—when soils are drying out) and between 10–15 days air frost in a month; confidence intervals after 15 days increase as the number of observations decrease. Temperature change over one week is not significant while temperature change over two weeks and SMD change over one and two weeks all show a small variation; however, the confidence intervals are large suggesting minimal effect on AC pipe failures. For the soil variables a higher number of AC pipe failures are observed in high soil pH (>11), in alluvial peats and clays with a large shrink swell potential of 12–15% volumetric expansion when drained to a depth of 2 m. Hydrologically impermeable (soft) soils (hydrology of soil type (HOST) category 8) comprising cover loam, clay with flints or plateau drift and loamy drift shows a higher number of AC pipe failures.

Steel and Ductile Iron

The results of the GAM PDPs for SDI pipes are presented in Fig. 3. Temperature, SMD, SMD change over one and two weeks are not significant in the model. Observing the PDPs, the variables revealing the largest effect on SDI pipe failures include days air frost and pipe diameter. As the number of days air frost in a month increases the SDI pipe failures increase, and there is an apparent increase in SDI pipe failures for smaller diameter pipes

Table 1. Concurvity test results.

	Categorical variables	Temperature	Temperature change over one week	Temperature change over two weeks	Days air frost	SMD	SMD change over one week	SMD change over two weeks	pH
Categorical variables	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Temperature		1.00	0.02	0.04	0.32	0.13	0.06	0.06	0.00
Temperature change over one week			1.00	0.10	0.02	0.01	0.02	0.01	0.00
Temperature change over two weeks				1.00	0.02	0.01	0.02	0.02	0.00
Days air frost					1.00	0.11	0.05	0.05	0.00
SMD						1.00	0.09	0.11	0.00
SMD change over one week							1.00	0.27	0.00
SMD change over two weeks								1.00	0.00
pH									1.00

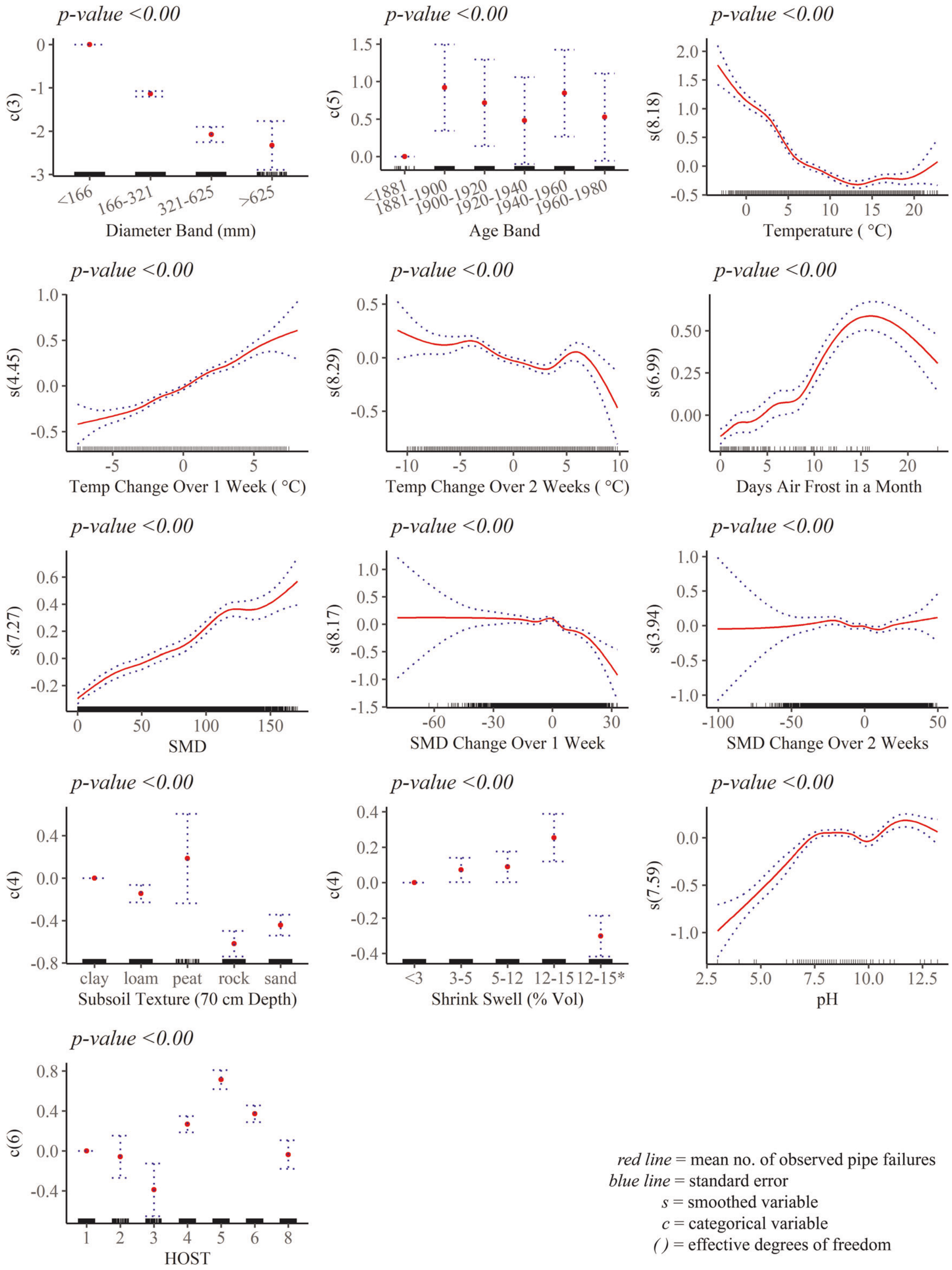

Fig. 1 Generalized additive model partial dependence plots for iron pipes. Each plot shows a covariate and their partial dependence on pipe failures in the context of the model. The y axis shows the mean number of observed failures and the x axis the covariate interval. The blue line represents the 95% confidence interval.

(small pipes < 166 mm). Pipe age indicated a fall in pipe failures for newer pipe installations; however, the confidence interval is wide. Observing shrink swell, the highest number of failures is noted in soils with low and medium shrink swell potential suggesting that soil movement has limited affect on SDI pipe failures. Peat and clay subsoils show the highest number of SDI failures, although the confidence interval for peat is high suggesting too few observations for this soil category. There are a higher number of SDI pipe failures in soils which are unconsolidated and macroporous (HOST category 5), these soils comprise mudstones, soft massive clays, very soft clays and loams. The remaining variables showed a limited effect on the number of pipe failures.

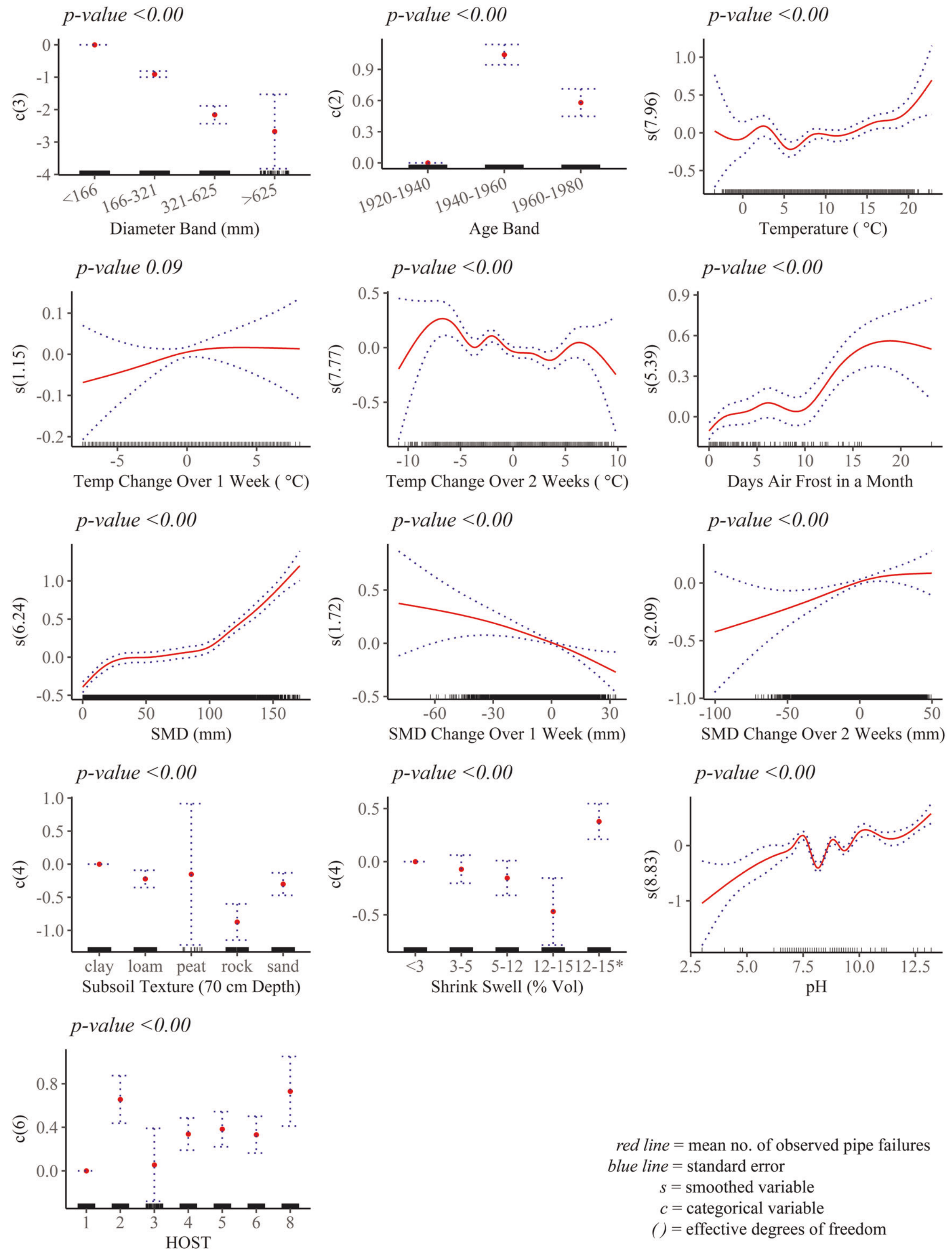

Fig. 2 Generalized additive model partial dependence plots for asbestos cement pipes. Each plot shows a covariate and their partial dependence on pipe failures in the context of the model. The y axis shows the mean number of observed failures and the x axis the covariate interval. The blue line represents the 95% confidence interval.

Polyvinyl chloride

The GAM PDPs for PVC pipes are presented in Fig. 4. The difference between the diameter bands shows a large uncertainty in the largest pipes >625 mm due to a small number of observations and shows no obvious variation between other diameter bands to suggest there is any real effect on PVC pipe failures. Pipe age shows little effect on pipe failures evidenced by the small variation between categories. SMD change over two weeks and temperature change over one week shows no significance in the model, and looking at the PDPs there is also a limited effect from temperature change over two weeks and SMD change over one week, both revealing little variation in PVC pipe failures and high confidence intervals. Observing soil pH, there is a minor increase in failures for more acidic soils; however, the confidence interval increases and

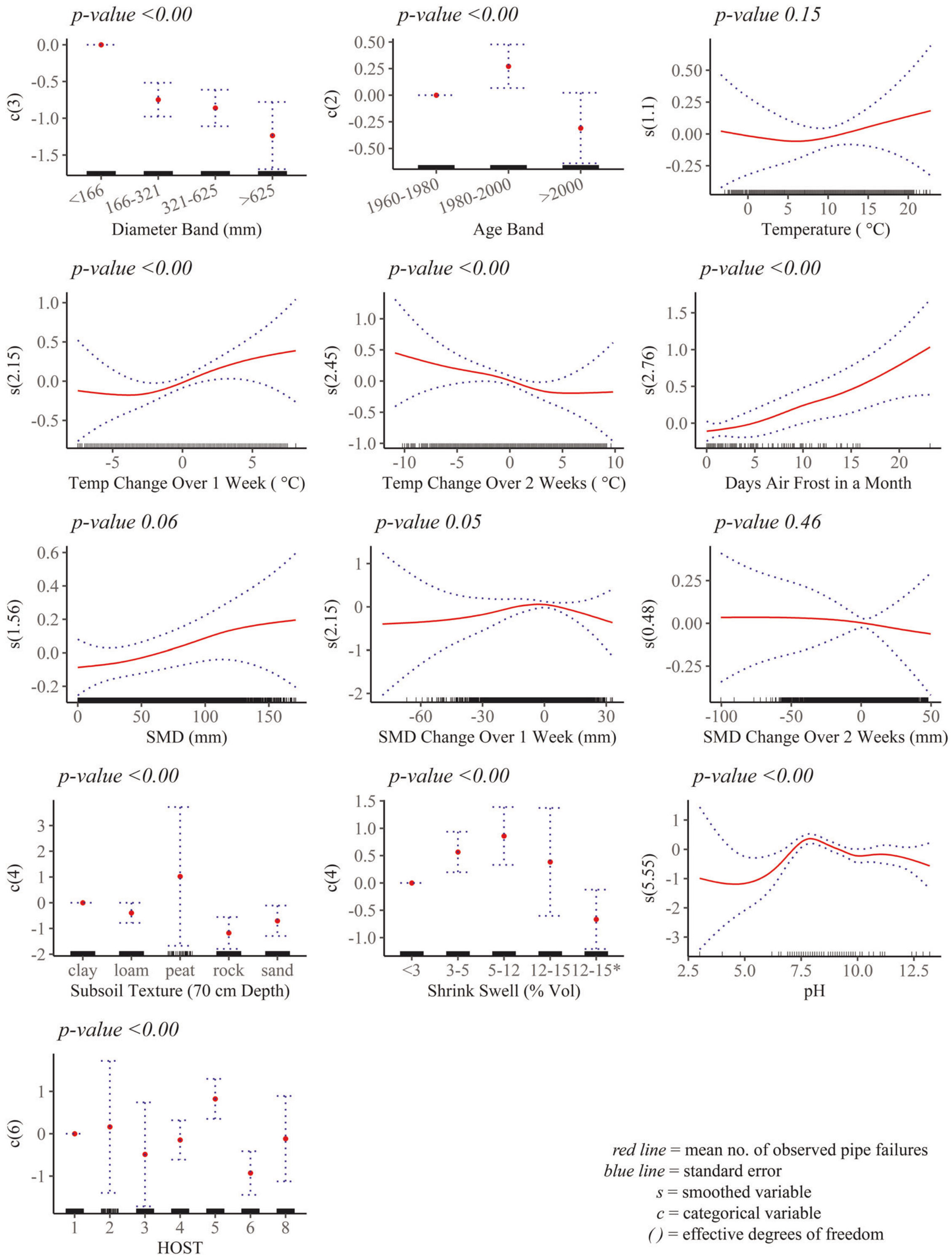

Fig. 3 Generalized additive model partial dependence plots for steel and ductile iron pipes. Each plot shows a covariate and their partial dependence on pipe failures in the context of the model. The y axis shows the mean number of observed failures and the x axis the covariate interval. The blue line represents the 95% confidence interval.

the number of observations is limited, therefore it is considered that highly acidic or alkaline soils has little effect on PVC pipe. Weather variables with the most effect on PVC failures include temperature and days air frost. Considering temperature, the number of PVC pipe failures increases as the temperature increases, and the same can be seen for days air frost in a month where the number of failures increases as the number of days frost in a month increases. Considering subsoil type, peat has the highest effect on PVC pipe failures, although the confidence interval is large due to a small number of observations. Loam and sand dominant soils show more certainty and have a minor effect on higher pipe failures. Soils with weakly consolidated macroporous soils (HOST category 2) comprising chalk and chalk rubble are also associated with the highest number of PVC failures.

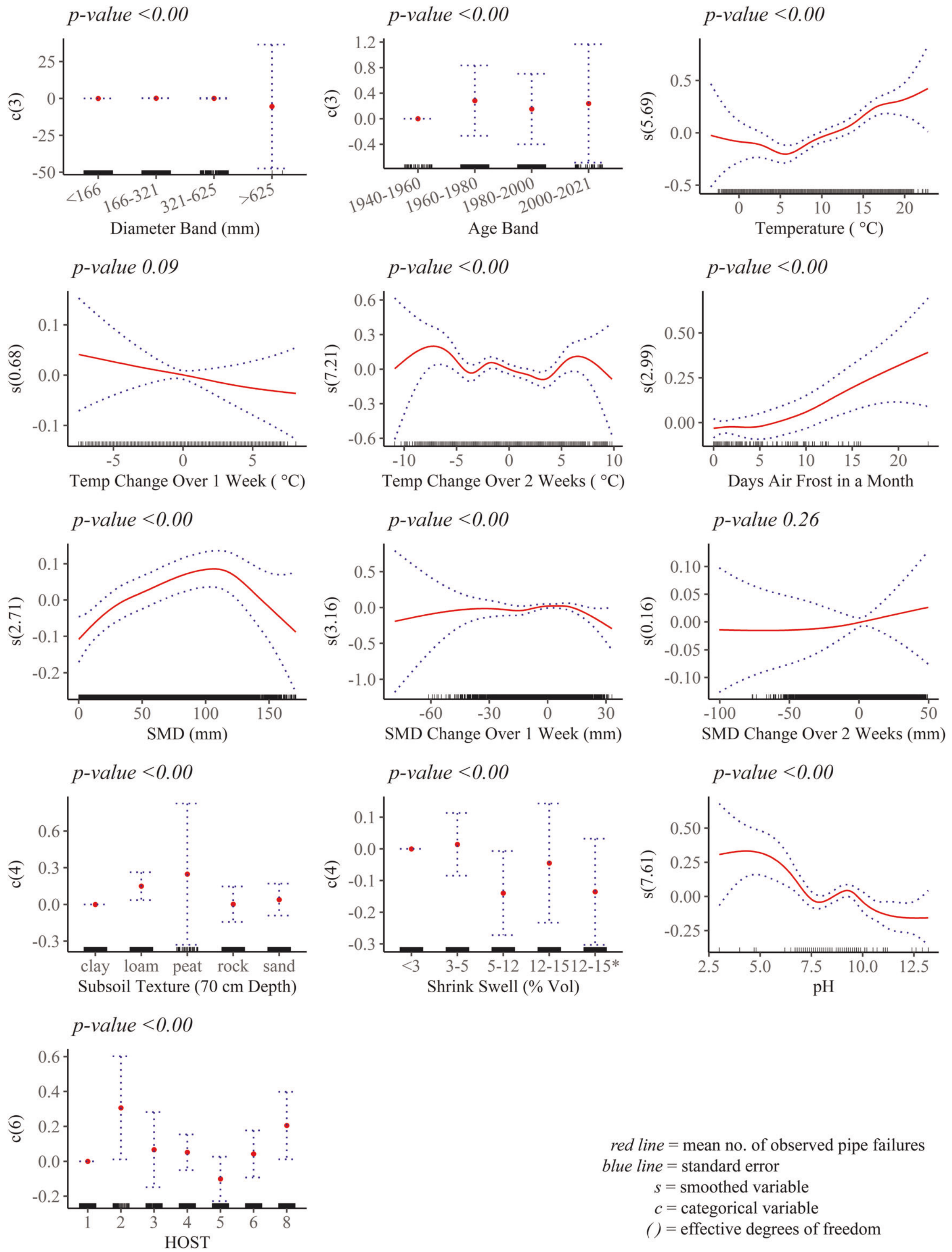

Fig. 4 Generalized additive model partial dependence plots for polyvinyl chloride pipes. Each plot shows a covariate and their partial dependence on pipe failures in the context of the model. The y axis shows the mean number of observed failures and the x axis the covariate interval. The blue line represents the 95% confidence interval.

Polyethylene

Figure 5 shows the GAM PDPs for PE and reveals that temperature and days air frost have the most effect on the number of PE pipe failures. Pipe failures increase for both low and high temperatures although confidence intervals increase for both high and low temperatures, while failures increase with more days air frost in a month up to 15 days, where the confidence intervals become high as the number of observations reduce. Temperature change over one and two weeks, and SMD and SMD change over one and two weeks show no significance (*p*-value > 0.01), and there is no obvious variation observed in the number of PE pipe failures. Soil pH, whilst narrowly significant does not indicate any obvious

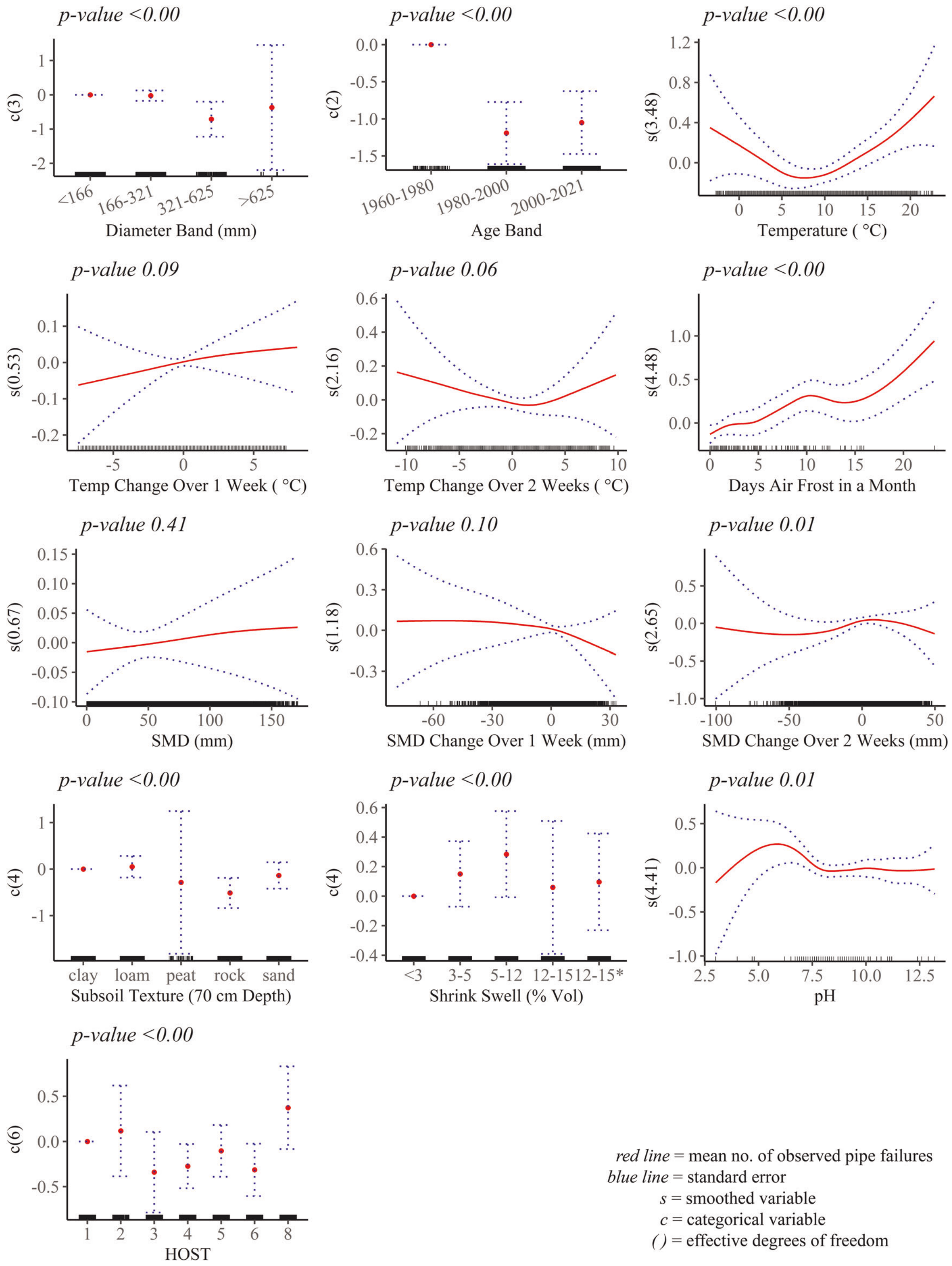

Fig. 5 Generalized additive model partial dependence plots for polyethylene pipes. Each plot shows a covariate and their partial dependence on pipe failures in the context of the model. The y axis shows the mean number of observed failures and the x axis the covariate interval. The blue line represents the 95% confidence interval.

pattern with PE pipe failure and alkaline or acidic soils. Diameter band also shows little variation between sizes but does suggest PE pipes of <321 mm have the highest number of failures. Age band indicates older PE pipe installed between 1960 and 1981 are associated with a higher number of failures. Considering soil characteristics, soil shrink swell potential and subsoil type revealed little variation and therefore only a small effect on the number of pipe failures, while hydrologically impermeable (soft) soils (HOST category 8) comprising cover loam, clay with flints or plateau drift and loamy drift showed a higher number of pipe failures.

DISCUSSION

Iron

Temperature, days air frost, SMD, peat and clay soils and pipes with a diameter <166 mm are the most influential covariates on pipe failures. In interpreting the results, iron pipe failures increase during winter when the temperature is cold, being particularly vulnerable to temperatures below 3 °C. This is attributed to the effects of air frost, and the subsequent cooling of soils adding additional stress on pipes[8,9,19]. Cooling of soils is also exacerbated by the presence of water and therefore, has more effect on soils with high water retention capacity. This is reflected in these data where the highest number of pipe failures for iron is observed in clay and peat soils, in line with previous studies[8]. Peat and clay soils typically retain more moisture after rainfall due to small pores and a slow permeability of <0.1 m day^{-1}. Therefore, higher rainfall during October and November is retained by these soils remaining saturated during the winter. As soils take on water so the thermal conductivity increases, ceasing when the soil is saturated (the more water in the soil the colder the soils can become). However, when moisture turns to ice during low temperatures and prolonged air frost, the thermal conductivity increases once again allowing soils to become colder[20]. Further to this soil cooling, the water retained in the soil pores freezes and increases in volume by ~9% resulting in soil movement[21]. Frost fronts can also develop as a result of prolonged cold temperatures and can cause soil-related ground movement. However, in the UK, where frost typically lasts less than two weeks, frost fronts develop near the soil surface, and as a result exert only limited consequence on pipes, which are typically buried at an average depth of 85 cm. Occasionally longer periods of frost and consequent frost penetration to greater depths can lead to a higher number of pipe failures[22].

The higher number of iron pipe failures observed in small diameter pipes <166 mm, can be attributed to the pipe walls being thinner, meaning they have less resilience to ground movement and corrosion and are therefore more likely to fail. Joints in small diameter pipes have also been reported to be less reliable than in larger diameter pipes[7]. There are also a significantly higher number of small diameter pipes compared to other pipe diameter sizes in the network.

Asbestos cement

AC pipe failures are higher when temperature and SMD is high, being typically associated with the summer months. This is due to the effects of ground movement of highly expandable clay soils, as a result of associated clay minerals (i.e. smectite, montmorillonite and vermiculite)[23] which are highly responsive to soil water content[24]; similar to findings in the Netherlands and Australia[7,8]. A high number of AC pipe failures is especially seen in alluvial clay and peat soils that have a high shrink swell potential of 12–15% volumetric expansion or greater if drained to a 2 m depth. These types of conditions would require a high evapotranspiration rate to realise the potential shrink swell hazard which might be associated with a period of drought. In addition to soil shrink swell potential in clay soils, it is possible that pipe failures are also associated with concrete corrosion, since pipe failures are slightly higher in highly alkaline soils (pH > 8.5), although the effect is small. Other proxies for soil corrosion could potentially be used to determine the effects of corrosion on AC pipes. A small increase in pipe failures is associated with low temperatures between 5 °C and 0 °C, coupled with a near linear increase in pipe failures with an increase in days air frost in a month. This again highlights the influence of prolonged periods of frost on pipe failure as found in comparable literature that reported higher joint failures during winter for both AC and PVC pipes[25]. The effects of pipe age suggests a peak in pipe failures between 1940 and 1960.

Pipe age is significant for AC and fewer failures are expected in new pipes. AC pipe was seen as a modern material during the 1940–1960s and was prolifically used. During this period 81% of AC pipes in the network were installed.

Steel and Ductile Iron

For SDI pipes, the length of pipe in the network and number of pipe failures is considerably lower than other materials due to the material's cost. Days air frost and diameter size showed the strongest partial effects on pipe failures, and suggests that SDI pipes fail as a result of the influence of frost on ground conditions, this has been discussed in detail for iron pipes (see the section "Iron"). SDI pipe failures are highest during the winter months, so the results here are not surprising. SDI pipe show a small increase in pipe failures during the summer months[11]. Observing the results to understand this trend, we look to see if soil shrink swell could be responsible. Despite the highest number of SDI pipe failures being observed in clay soil, there is little evidence to suggest soil shrink swell is the cause, since SDI pipe failures reduce as the potential for volumetric expansion increases. Therefore, other factors not included in this study may explain this seasonal trend observed, such as water usage or pipe pressure.

Polyvinyl chloride

The PDPs suggest that a limited number of environmental variables effect PVC pipe failures. PVC failures do increase as temperature increases, with the highest mean failure rate observed during the hottest temperatures, similar to AC pipes[6]; however above 16 °C the uncertainty bounds of these data also increase. PVC shows no relationship with soil shrink swell, with little variation in pipe failures from low to high classes, and is likely to be related to the plasticity of the material and the flexibility of the push fit joint most commonly used in PVC pipes. PVC pipe failures were highest in peat subsoils (noting however the uncertainty bounds of peat were very high) and in loam soil, but the effects of soil are small. It is also worth noting that flexible fitting joints such as the push fit joints typically used on PVC pipes can absorb soil displacement to a certain degree, which reduces stress on PVC pipes in soils acceptable to ground movement[25]. Considering this and the small effects of SMD up to ~120 mm, we suggest a small effect from drying soils, but this does not explain the high summer failures for PVC which are likely to be associated with other contributing factors. Previous studies have suggested that internal pressure from increased water demand during the summer months may be the reason for PVC pipe failures[8]. This is reported as particularly detrimental in push-fit joints since they do not provide restraint from thrust from internal pressure forces, resulting in axial pull out or disconnection[26,27]. In these data we did not have access to either pressure data or failure type, therefore further investigation is required to support the literature in this area and understand the main failure mechanism. Failures may also be associated with soil washout potential, where an existing pipe failure can wash soil from under another, causing the pipe to bridge, sag and eventually fail[28]. Considering the effects of days air frost, PVC shows a degree of vulnerability to long periods of frost, similar to iron pipes and as reported by existing literature[25]. However, considering the uncertainty interval the effect may be minor.

Polyethylene

PE pipes show little variation in number of failures in the PDPs. Temperature shows a higher number of failures at low temperatures below 5 °C, which is likely to be related to the effects of frost on soils since failures increase as the number of days air frost increases. However, the wide confidence intervals suggest the effects are small. The winter failures may be associated with joint failure as per PVC and AC[29] since the material's ability to withstand thermal expansion and contraction (plasticity) makes it more resilient to ground movement. However, this could only occur on joints that were not properly fused together, since PE pipes connected correctly through electrofusion rarely fail. Due to the plasticity of PE pipes, and limited partial effects in different soils and shrink swell, the increase in pipe failures during high temperatures are likely to be related to other factors such as increased water demand. Age of PE pipe shows that early pipes installed between 1960 and 1981 are associated with a higher number of failures (Fig. 5). This could be due to early PE material being a different grade of PE or improvements in joining pipes shifting from plate welding and clamping to electrofusion[11].

Concluding remarks

Using a multivariate GAM approach with PDPs provided a flexible and useful tool for revealing insight into complex relationships and structure in these data. The findings of this study contribute to existing literature and can be used to aid a further understanding of environmental interactions and how they affect pipe failures for different material types. A natural extension of these findings would be to build future statistical models with knowledge of which environmental variables are likely to explain pipe failures for each material type. The results found here for an entire UK network are similar to comparable studies from other countries (Netherlands[6,8] Austria[9] and Australia[7]), and will be useful to water utility companies, who could use the information to understand failures in similar network conditions. The prominent findings of the study include:

- Some of the continuous variables smoothed by the GAM displayed low effective degrees of freedom. Where this occurred the variables may provide better correlation as linear predictors.
- Diameter band is important for all materials, except PVC. Higher failure rates occurred in pipes with a diameter of <166 mm.
- Pipe age is important for all pipe materials except iron and PVC.
- Severe temperatures have a significant influence on pipe failures for all materials, however the influence of temperature strongly depends on the type of pipe material.
- Soil pH, considered a marker for corrosion, has far less influence than was initially anticipated. Other soil corrosion proxies such as a soil corrosion index may yield better results.
- A rapid change in both temperature and SMD as determined in this study has a weak effect on pipe failures for all materials.
- Iron and SDI pipes fail mainly in the winter, during temperatures <5 °C and long periods of air frost. Clay and peat soils with high moisture retention show the highest number of failures during the winter.
- AC pipes mainly fail during the summer as a result of increasing temperatures, high SMD and subsequent shrink swell in highly shrinkable clay and peat dominant soils causing soil movement.
- PVC pipe failures increase during the summer and are affected by temperature and SMD. However, ground movement does not appear to be the only cause. Therefore, other factors not explored in this study must be able to explain PVC summer failures.
- PE pipes exhibit a higher number of failures in severely cold temperatures with prolonged air frost.

Understanding the effects of different environmental factors on pipe failure is an important step when building data-driven, statistical infrastructure models. This understanding can help to reduce common confusion of correlation and causation effects and can help in the selection of appropriate variables and their derivatives. Such information can lead to more accurate and robust models, which in turn can be used by utilities to make informed network management decisions, aiming at reducing water lost through network leaks and helping facilitate future water demand.

METHODS

Data collection

Historical data collected from 2005 to 2018 have been obtained from three sources: network distribution data from a water company, weather data from the UK Met Office and the national soil map-related Natural Perils Directory and LandIS soils data and maps from Cranfield University[30].

The network distribution data are for an entire drinking water network which covers a geographical area of 27,476 km^2 in the UK, comprising of ~40,000 km of pipeline. Datasets representing the network contain information regarding each asset including the location of the drinking water distribution network, and the historic burst locations. The different infrastructure variables included in this dataset include length of pipe, diameter size, year of installation, and age of pipe. Two types of bursts are recorded by the infrastructure operator, namely reactive and proactive. Reactive bursts are those responded to once they are reported (by the infrastructure operator or the customer), while proactive bursts are those actively investigated by the operations team or which can be found during a campaign or 'sweep' of an area; thus, the date of repair is temporally unrelated to the date of failure. Only reactive bursts have been included in this analysis, since proactive bursts can manifest for long periods of time before they are finally found and repaired, which makes it hard to ascertain the original date of failure or to link the failure to dynamic temporal environmental variables, so potentially distorting the final results. A summary of the drinking water network used in our analysis is provided in Table 2.

Cranfield University's Natural Perils Directory and LandIS datasets have been used to determine the influence of soil on water infrastructure[30]. These data include ground movement (shrink swell) and pH presented as 1:250,000 maps based on field data collected between 1939 and 1987. LandIS provides the soil data including soil substrate class and HOST[31] which classifies soil according to its hydrological behaviour. The authors have broadly categorised the 29 UK wide HOST classes into 10 substrate hydrology classes, of which only 8 occurred within the study area (see Supplementary Table 2).

The Met Office historical weather data have been collected from the Met Office Rainfall and Evaporation Calculation System (MORECS) and the summary datasets website[32]. MORECS data provide weekly estimates of SMD, hydrologically effective runoff (rainfall) and temperature, expressed on a gridded basis with a spatial resolution of 40 × 40 km.

Data assimilation

Spatial data processing were undertaken using ArcGIS (version 10.6)[33] while data assimilation, processing, modelling and visualisation were completed using R software (version 3.2.3)[34]. Using ArcGIS, reactive pipe failure records collected onsite were relocated (snapped) to the nearest matching pipes using a distance of 3 m, based on the nearest pipe with equivalent diameter size and material type to those recorded onsite.

Table 2. Summary of drinking water network used in the analysis.

Material	Installation year	Total network length (km)[a]	Failure rate/ km/year
Iron	1881 to 1921	11,382	0.27
Asbestos cement (AC)	1920 to 1941	7090	0.17
Steel and ductile iron (SDI)	1960 to present	1824	0.05
Polyvinyl chloride (PVC)	1960 to 2001	5854	0.17
Polyethylene (PE)	1981 to present	12,274	0.03
Total	1881 to present	38,424	0.13

[a] Length of network currently operational by end of December 2018.

Where pipe failures did not match a pipe section located within this 3 m distance, the process was repeated again sequentially up to a distance of 1 km until a match was found. The failure data were then joined to the pipe data using geographical reference. Pipes were attributed through spatial intersection with shrink swell, pH and HOST. The MORECS[35] and summary weather data metrics were derived to comprise variants of the original data. Pipe diameter and age, which were classified into appropriate bands (Tables 3 and 4).

These data were split into cohorts of pipes based on the following characteristics: material, age, diameter, 1 km grid, soil type (sand, clay, peat, silt and soft). Dividing, or sorting the pipes into homogeneous groups was found useful for analysis as it reduces the computation time required for big data and reduces data uncertainty; a common approach adopted for statistical modelling of pipes[36,37]. These data were then summarised by week to join with the weekly weather data.

Table 3. Classification bands for pipe age.

Age band	Range
1	<1880
2	1880 to <1900
3	1900 to <1920
4	1920 to <1940
5	1940 to <1960
6	1960 to <1980
7	1980 to <2000
8	2000 to <2021

Table 4. Classification bands for pipe diameter.

Diameter band	Range
1	<166 mm
2	166 to <321 mm
3	321 to <626 mm
4	≥626 mm

The environmental variables considered for analysis are outlined in Table 5 along with the thresholds and units considered. The environmental variables have been selected based on a review of relevant literature undertaken[11] that identified the major environmental factors and their influence on pipe failure for different materials.

Statistical analysis

Data exploration was undertaken to ascertain the effects of environmental conditions of pipe failure used by modelling the variables against the number of pipe bursts (total number of bursts divided by the pipe length) using a GAM and viewing the results of the PDPs. GAMs are similar to Generalized Linear Models (GLM) but differ by relaxing the linear assumption, potentially revealing non-linear relationships and important structure in these data that would otherwise be missed[38]. To this extent, a GAM can show either linear, monotonic or more complex relationships, depending on the way each variable responds to changes in the dependent variables. The GAM performs in this way by extending a GLM to include a smoothing basis function that can measure arbitrarily non-parametric relationships. Categorical data are treated as a linear term without smoothing[39]. This semi-parametric approach makes GAMs very flexible, easily accommodating different types of data[38]. It should be noted that PDPs are limited in the sense that they only apply in the context of the model used. PDPs may also be misleading in the presence of variable dependence (known as concurvity in GAMs).

The computational methods were implemented from the cran repository 'mgcv' package. For the GAM modelling we assume a Poisson distribution since pipe failures are count data. The notation for the GAM smoothing in a Poisson model is as follows[40,41].

$$g(E(y_i)) = \beta_0 + \beta_j x_{ji} + \cdots + f_k(x_{ki})\cdots,$$

where $(E(y_i))$ is one of n observations of the response variable, g the Poisson distributed exponential family with the log link function, β_0 is the mean number of observed pipe failures, β_j is the linear term of some predictor covariate x_i and f_k the smoothing term of some non-parametric predictor covariate x_k. to overcome some issues of working with big data[42]. From the 'mgcv' package we use 'bam', a less memory intensive version of 'gam'[43]. A log offset for pipe length was included as an explanatory variable to correct the difference in pipe length. For the smoothing basis function, we use the penalised cubic regression spline. This is beneficial since it lowers computation cost and avoids overfitting by applying a

Table 5. Variables considered for statistical analysis, description and units.

Variable	Description	Units
Pipe failures	Number of reactive bursts	No.
Length of pipe	Total length of each pipe asset split by asset number	m
Pipe age	Pipe age categorised into eight categories (see Table 3)	Years
Pipe diameter	Pipe diameter categorised into four categories (see Table 4)	mm
Temperature	Temperature change in 1 °C increments	°C
	Temperature change in 1 °C over 1 week. (weekly mean minus the next weekly mean)	
	Temperature change in 1 °C over 2 weeks. (fortnightly mean minus the next fortnightly mean)	
Days air frost	Total days air frost in a month	Days
Soil Moisture Deficit (SMD)	SMD levels	mm
	SMD change over 1 week	
	SMD change over 2 weeks	
Soil texture	Subsoil texture based the dominant class in the soil association at a depth of 70 cm. Soils broadly categorised into clay, loam, peat, rock and sand	Categories
Shrink swell	Six categorical levels of soil shrink swell potential with a % volumetric expansion <3, 3–5, 5–12, 12–15 and 12–15*	Categories[a]
pH	Dominant average pH value in the soil based on the soil classification	pH value
Hydrology of soil type	HOST classes are based on substrate hydrogeology flow mechanism and is split into 8 sub classes	Categories[b]

[a] Supplementary Table 1.
[b] Supplementary Table 2.
*when soils are drained to a depth of 2 m.

smoothing penalty when estimating the coefficient, generalising the smoothers by shrinking them towards zero. This penalisation approach also helps to generalise the model which can reduce the effects of concurvity and when using 'select = TRUE' can remove variables from the model if they add no value[41]. We leave the number of knots as default unless residuals are found to be significant, in which case we adjust upwards accordingly. The smoothing parameter estimation was restricted maximum likelihood ('REML'), typically used for smooth components viewed as random effects[38]. The notation for the penalised cubic regression is represented for two variables as follows[41]:

$$f_1(x_1) = \sum_{j-1}^{q_1} \beta_j b_{1j}(x_1)\ f_2(x_2) = \sum_{j-1}^{q_2} \beta_{j+q_1} b_{2j}(x_2),$$

where q_j represents the penalty matrix, β_j the unknown coefficient to be estimated and $b_j(x_j)$ and the cubic regression spline basis function for f_j of a non-parametric predictor covariate x_j[41].

The occurrence of variable dependence (known as concurvity in GAM models) can result in poor parameter estimation and increased confidence intervals, leading to an increased risk of a false statistically significant effect. The use of penalized cubic regression is an approach that helps reduce the effects of concurvity[39], nonetheless it is important to check these data. Firstly, we check to ensure that the model has converged, therefore no major errors have occurred. We then use the 'concurvity()' function from the 'mgcv' package which measures how a smoothed variable can be approximated by another. All categorical variables are grouped together and measured against each smoothed variable resulting in a single value for concurvity. The concurvity output metric presents three scenarios, worst, observed and estimated. The estimate is the most reliable measure for concurvity which returns a value between zero and one, where 0 = no concurvity and 1 = no identifiability between the variables[42]. There is no universal criteria for concurvity, but one study on the effects of concurvity suggests values >0.5 starts to introduce noticeable errors[44]; we adopt this value as the cut off.

The final results are presented as PDPs using the cran 'mgcViz' package[18]. PDPs are a useful model agnostic tool used to perform a graphical exploratory evaluation of the variable effects by showing the change in the mean predicted number of pipe failures as the variable interval changes over its distribution (mean centred since the smoothed variables must sum to zero in a GAM)[45], accompanied by the 95% confidence intervals. The estimated p-value (<0.01) will be used to determine the significance of the variable effect on pipe failures (Supplementary R script).

DATA AVAILABILITY

The network data are not publicly available due to their commercial nature which the water company has not made publicly available. The weather data are available under licence from www.metoffice.gov.uk. The soils data used are available under licence from www.landis.org.uk.

Received: 5 November 2019; Accepted: 4 June 2020;
Published online: 01 July 2020

REFERENCES

1. Environment Agency. Environment Agency calls for action on water efficiency - GOV.UK. https://www.gov.uk/government/news/environment-agency-calls-for-action-on-water-efficiency (2018).
2. Farewell, T. S., Jude, S. & Pritchard, O. G. How the impacts of burst water mains are influenced by soil sand content. *Nat. Hazards Earth Syst. Sci.* **18**, 2951–2968 (2018).
3. Pritchard, O. G., Hallett, S. H. & Farewell, T. S. Soil impacts on UK infrastructure: current and future climate. *Proc. Inst. Civ. Eng. - Eng. Sustain.* **167**, 170–184 (2014).
4. Scheidegger, A., Leitão, J. P. & Scholten, L. Statistical failure models for water distribution pipes – A review from a unified perspective. *Water Res.* **83**, 237–247 (2015).
5. Pelletier, G., Mailhot, A. & Villeneuve, J. P. Modeling Water Pipe Breaks—Three Case Studies. *J. Water Resour. Plan. Manag.* **129**, 115–123 (2003).
6. Wols, B. A. & van Thienen, P. Impact of weather conditions on pipe failure: a statistical analysis. *J. Water Supply Res. Technol.* **63**, 212–223 (2014).
7. Gould, S. J. F., Boulaire, F. A., Burn, S., Zhao, X. L. & Kodikara, J. K. Seasonal factors influencing the failure of buried water reticulation pipes. *Water Sci. Technol.* **63**, 2692–2699 (2011).
8. Wols, B. A., Vogelaar, A., Moerman, A. & Raterman, B. Effects of weather conditions on drinking water distribution pipe failures in the Netherlands. *Water Sci. Technol. Water Supply* **19**, 404–416 (2019).
9. Fuchs-Hanusch, D., Friedl, F., Scheucher, R., Kogseder, B. & Muschalla, D. Effect of seasonal climatic variance on water main failure frequencies in moderate climate regions. *Water Sci. Technol. Water Supply* **13**, 435–446 (2013).
10. Kakoudakis, K., Behzadian, K., Farmani, R. & Butler, D. Pipeline failure prediction in water distribution networks using evolutionary polynomial regression combined with K-means clustering. *Urban Water J.* **14**, 737–742 (2017).
11. Barton, N. A., Farewell, T. S., Hallett, S. H. & Acland, T. F. Improving pipe failure predictions: factors affecting pipe failure in drinking water networks. *Water Res.* **164**, 114926 (2019).
12. Bruaset, S. & Sægrov, S. An analysis of the potential impact of climate change on the structural reliability of drinking water pipes in cold climate regions. *Water* **10**, 411 (2018).
13. Rajani, B., Kleiner, Y. & Sink, J.-E. Exploration of the relationship between water main breaks and temperature covariates. *Urban Water J.* **9**, 67–84 (2012).
14. Farewell, T. S., Hallett, S. H. & Truckell, I. G. Soil and climatic causes of water mains infrastructure bursts. *Natl. Soils Res. Ins.* (2012).
15. Heinze, G., Wallisch, C. & Dunkler, D. Variable selection – A review and recommendations for the practicing statistician. *Biometrical J.* **60**, 431–449 (2018).
16. Ratner, B. Variable selection methods in regression: ignorable problem, outing notable solution. *J. Targeting, Meas. Anal. Mark.* **18**, 65–75 (2010).
17. Marra, G. & Wood, S. N. Practical variable selection for generalized additive models. *Comput. Stat. Data Anal.* **55**, 2372–2387 (2011).
18. Fasiolo, M., Nedellec, R., Goude, Y. & Wood, S. N. Scalable visualization methods for modern generalized additive models. *J. Comput. Graph. Stat.* **29**, 78–86 (2019).
19. Kakoudakis, K., Farmani, R. & Butler, D. Pipeline failure prediction in water distribution networks using weather conditions as explanatory factors. *J. Hydroinformatics* **20**, 1191–1200 (2018).
20. Becker, B. R., Misra, A. & Fricke, B. A. Development of correlations for soil thermal conductivity. *Int. Commun. Heat Mass Transf.* **19**, 59–68 (1992).
21. Trickey, S. A., Moore, I. D. & Balkaya, M. Parametric study of frost-induced bending moments in buried cast iron water pipes. *Tunn. Undergr. Sp. Technol.* **51**, 291–300 (2016).
22. Farewell, T. S., Hallett, S. H., Hannam, J. A. & Jones, R. J. A. Soil impacts on national infrastructure in the UK. *Infrastructure Transitions Res. Consort.* 1–45 (2012).
23. Pritchard, O., Farwell, T. & Hallett, S. Soil-related geohazard assessment for climate-resilient UK infrastructure. *Cranf. Univ.* 1–400 (2015).
24. Pritchard, O. G., Hallett, S. H. & Farewell, T. S. Probabilistic soil moisture projections to assess Great Britain's future clay-related subsidence hazard. *Clim. Change* **133**, 635–650 (2015).
25. Wols, B. A. & van Thienen, P. Modelling the effect of climate change induced soil settling on drinking water distribution pipes. *Comput. Geotech.* **55**, 240–247 (2014).
26. Arsénio, A. M., Pieterse, I., Vreeburg, J., De Bont, R. & Rietveld, L. Failure mechanisms and condition assessment of PVC push-fit joints in drinking water networks. *J. Water Supply Res. Technol. - AQUA* **62**, 78–85 (2013).
27. Sundberg, C. Fundementals of Differential Settlement of Pipelines. In *Sessions of the Pipelines Conference, Phoenix, AZ, AUG 06–09, 2017* (eds Pridmore, A. B. & Geisbush, J.) 469–481 (Pipelines 2017. Planning and design: proceedings of sessions of the Pipelines 2017 Conference, August 6–9, 2017, Phoenix, Arizona, 2017).
28. Pritchard, O., Hallett, S. H. & Farewell, T. S. Soil Corrosivity in the UK – Impacts on Critical Infrastructure. *Ifrastructure Transitions Res. Consort*. 1–55 (2013).
29. Ruchti, G. F. *Water Pipeline Condition Assessment* (American Society of Civil Engineers (ASCE), 2017).
30. Hallett, S. H., Sakrabani, R., Keay, C. A. & Hannam, J. A. Developments in land information systems: examples demonstrating land resource management capabilities and options. *Soil Use Manag.* **33**, 514–529 (2017).
31. Boorman, D. B., Hollis, J. M. & Lilly, A. *Hydrology of soil Types: A Hydrologically Based Classification of the Soils of United Kingdom.* (Institute of Hydrology, 1995).
32. Met Office. East_Anglia Days of Air Frost. https://www.metoffice.gov.uk/pub/data/weather/uk/climate/datasets/AirFrost/ranked/East_Anglia.txt (2019).
33. ESRI. ArcGIS Online | Interactive Maps Connecting People, Locations & Data. https://www.arcgis.com/index.html (2019).
34. R Core Team. R: The R Project for Statistical Computing. https://www.r-project.org/ (2018).
35. Hough, M. N. & Jones, R. J. A. The United Kingdom Meteorological Office rainfall and evaporation calculation system: MORECS version 2.0-an overview. *Hydrol. Earth Syst. Sci.* **1**, 227–239 (1997).
36. Zhou, Y. *Deterioration and Optimal Rehabilitation Modelling for Urban Water Distribution Systems. Deterioration and Optimal Rehabilitation Modelling for Urban Water Distribution Systems.* Delft University of Technology. 1–249 (2018).
37. Rajani, B. B. & Kleiner, Y. Using limited data to assess future needs. *Am. Water Work. Assoc.* **91**, 47–61 (1999).

38. Wiley, M. & Wiley, J. F. GAMs. In *Advanced R Statistical Programming and Data Models* 165–224 (Apress, 2019).
39. Buja, A., Hastie, T. & Tibshirani, R. Linear smoothers and additive models. *Ann. Stat.* **17**, 453–510 (1989).
40. Hastie, T. & Tibshirani, R. Generalized additive models. *Stat. Sci.* **1**, 297–310 (1986).
41. Wood, S. N. & Augustin, N. H. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecol. Modell.* **157**, 157–177 (2002).
42. Wood, S. N. *Generalized Additive Models: An Introduction with R*. (Chapman and Hall/CRC, 2006).
43. Wood, S. N., Goude, Y. & Shaw, S. Generalized additive models for large data sets. *J. R. Stat. Soc. Ser. C Appl. Stat.* **64**, 139–155 (2015).
44. He, S. *Generalized Additive Models for Data With Concurvity: Statistical Issues and a Novel Model Fitting Approach.* University of Pittsburgh. 1–51 (2004).
45. Goldstein, A., Kapelner, A., Bleich, J. & Pitkin, E. Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. *J. Comput. Graph. Stat.* **24**, 44–65 (2015).

ACKNOWLEDGEMENTS

This work was supported by the UK Natural Environment Research Council [NERC Ref: NE/M009009/1] and Anglian Water plc. The authors are grateful for their support.

AUTHOR CONTRIBUTIONS

N.B.: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Writing—Original Draft, Visualization, Data Curation. T.F.: Project Administration, Validation, Formal Analysis, Resources, Writing—Review and Editing, Supervision, Data Curation, Funding Acquisition. S.H.: Writing—Review and Editing, Supervision, Funding Acquisition.

COMPETING INTERESTS

The authors declare no conflict of interest.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41545-020-0077-3.

Correspondence and requests for materials should be addressed to T.S.F.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.