AQUA
Water Infrastructure, Ecosystems and Society

IWA PUBLISHING

 AQUA — Water Infrastructure, Ecosystems and Society Vol 72 No 1, 111 doi: 10.2166/aqua.2022.194

A model for evaluating water distribution system capacity as a function of the total pipeline length

Carlo Loubser *, Frans Grotepass, Jessica May Winter and Heinz Erasmus Jacobs

Department of Civil Engineering, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa
*Corresponding author. E-mail: carloloubser@gmail.com

CL, 0000-0001-9705-0298; HEJ, 0000-0002-7360-6375

ABSTRACT

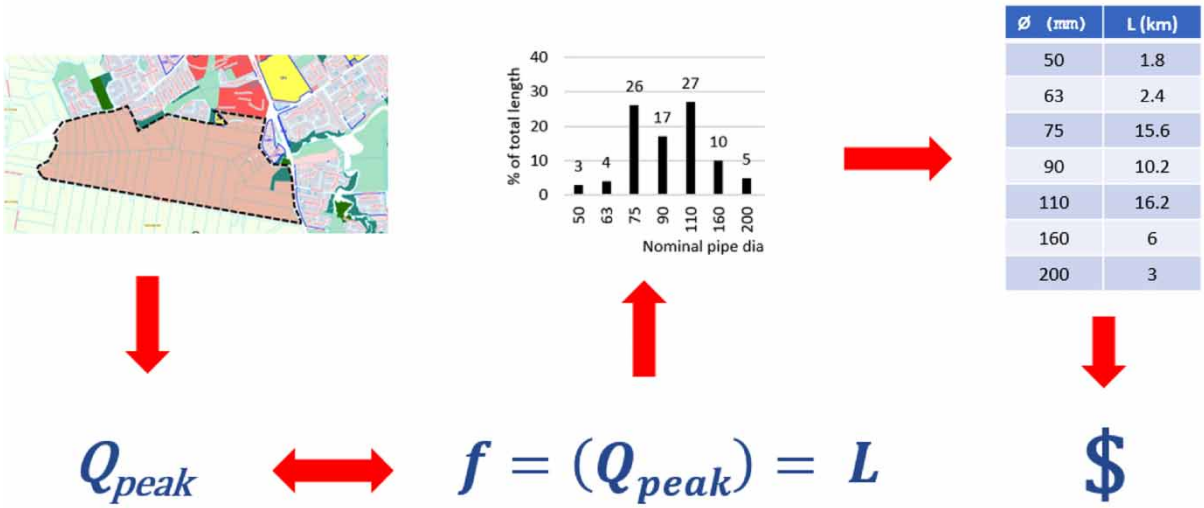
Planners are often faced with the challenge to provide crude estimates of water distribution system (WDS) infrastructure capacity and associated costs in the early phases of greenfield developments. This study investigated the relationship between the physical and hydraulic characteristics of a WDS and the corresponding serviced area. Five physical parameters (a) and two hydraulic parameters (b) describing the serviced area were identified for analysis, namely (a) total pipeline length, land area, area shape factor, terrain index, reservoir distance from area centroid and (b) peak flow rate and average static system pressure. Multiple linear regression was performed on the data. A model was compiled linking the total pipeline length of a WDS to the peak flow rate. The model is applicable to predominantly residential service zones larger than 80 hectares with a peak hour flow rate of <450 L/s. The model enables the prediction of the potable water distribution pipe infrastructure required for future development areas in the absence of basic planning information, such as cadastral layouts. Alternatively, the model can estimate the potential maximum peak flow rate that can be supplied, if the total pipeline length is known.

Key words: capacity, diameters, intermittent, pipeline length, regression, water distribution system

HIGHLIGHTS

- A large dataset of water distribution systems from South Africa was analysed.
- Seven parameters were evaluated using multiple linear regression.
- A model was compiled linking the total pipeline length of a water distribution system to the peak flow rate.
- Total pipeline length was segregated into diameter distributions.
- Prediction of the water distribution pipe infrastructure required for future development areas.

GRAPHICAL ABSTRACT



INTRODUCTION

Background

A water distribution system (WDS) consists of a network of pressurised pipes of varying diameters. These networks encompass multiple parameters such as pipeline diameter and associated lengths, internal roughness coefficients, available supply pressures, consumers' demand and the related peak hour flow rate. In order to manage all these variables, engineers require sophisticated computer programmes for hydraulic modelling. Master plans for WDSs are typically compiled by experienced modellers relying on accurate information. Engineers and city planners often have the need to crudely estimate the pipeline lengths and associated pipeline diameters in the early planning phases of future development areas – even before a cadastral layout is available.

Several tools have been proposed to predict the required water and sewer pipeline infrastructure in developed areas. Dames & Moore (1978) conducted a national survey of 455 sewer construction projects in the United States of America. One of the fundamental outcomes of the work by Dames & Moore (1978) was a table linking sewer pipe lengths and diameters to population size groups.

Sitzenfrei *et al.* (2010a) developed a virtual infrastructure benchmarking tool (VIBe), which algorithmically generates complex virtual case studies at a city scale for urban water systems, including sewer systems and water distribution systems. The parameters of the virtual case studies are stochastically varied in ranges extracted from real-world case studies and literature to cover a broad range of possible system properties (Sitzenfrei *et al.* 2010a). A module was added to this algorithm allowing sewer infrastructure to be developed for each virtual urban case study (Urich *et al.* 2010). VIBe is therefore used to generate the input files for the sewer network building software.

The VIBe algorithm was further enhanced by adding functionality to dynamically model time-related impacts on the urban structure for example changing land use and population size. This enhanced version of VIBe was named 'Dynamic Virtual Infrastructure Benchmarking' or DynaVIBe (Sitzenfrei *et al.* 2010b). DynaVIBe could be a useful model to generate dynamic virtual sewer case studies for a selected modelling scenario. The model input parameters can be varied to perform a sensitivity analysis and generate a realistic envelope of the infrastructure requirements for an area and be used to model various future scenarios.

Venkatesh & Brattebø (2011) performed analyses on a dataset of pipeline lengths of 30 Norwegian municipalities, in order to illustrate that the relationship between pipeline length and the number of pipes by length class can be defined by the power law.

Kobayashi *et al.* (2011) developed a model to predict the distribution of water pipeline lengths based on road layouts in Japan. Kobayashi's model was intended to improve the accuracy of earthquake damage assessment.

Maurer *et al.* (2012) developed a generic model of length, diameter distribution and replacement costs for the sewer network in a settlement with a fixed area. The model catered for three classes of pipes, namely private connection pipes, secondary sewer pipelines and the sewer trunk main that connects the settlement to the rest of the network.

Pauliuk *et al.* (2014) provided a calibrated estimate of the total length, total mass of pipelines and the diameter distribution for sewer networks in cases where only area and population density of the settlement are known. A link was established between two planning parameters (urban density and settlement size) and the demand for pipes and materials in water and sewer network infrastructure.

Balaji *et al.* (2015) used data from completed sewerage schemes in 31 towns in India to perform regression analyses. Balaji *et al.* (2015) determined empirical equations relating the total installation cost of sewer networks (defined as material, equipment and labour costs for excavation, laying and jointing) to the serviced population size.

The South African Department of Water and Sanitation (DWS) developed a cost benchmark for water services (DWS 2016). The outcome was a document which provides typical unit costs of water services projects and individual infrastructure components. The relevant costs were derived from the DWS rural water supply projects completed after 1994 and from as-built project costs from various consulting engineering firms in South Africa.

Several researchers attempted to model sewer network layouts and optimise sewer network design. Turan *et al.* (2019) developed a graph theory-based methodology for sewer system optimisation. The proposed method generates a viable sewer network layout that contains all sewer links and satisfies the requirements of a sewer system by using graph theory, without any additional strategies required. Hesarkazzazi *et al.* (2022) proposed a graph theory-based framework for sewer system layout. In addition, a generic scheme for decentralised layouts in both steep and flat terrains was suggested. Duque

et al. (2022) proposed a spatial algorithm for generating simplified sewer networks which represent key characteristics of real systems, using basic topographic, demographic and urban characteristics. Three different pipe dimensioning approaches were compared and a balance between detail and computational efficiency was found. Moeini & Afshar (2018) used the ant colony optimisation algorithm in combination with nonlinear programming techniques for the optimal design of sewer networks. The ant colony optimisation algorithm was used to determine pipeline diameters, while nonlinear programming was used to determine the pipeline slopes.

Winter *et al.* (2022) used multiple linear regression to estimate the total sewer pipeline length for a service zone using basic service zone characteristics. Pipeline diameter distributions were developed for disaggregating the total pipeline length into lengths per diameter. In addition, the number of manholes required along a length of the pipeline for different types of service zones was quantified.

Estimating the required pipeline infrastructure of water and sewer networks based on limited information has been attempted before. Tools for the automatic generation of water and sewer network infrastructure have the potential for estimating infrastructure and for the high-level costing thereof. The benefit of direct costing methods is that minimal information is required to develop cost estimates. Tools that enable customisation for specific conditions such as the DWS cost benchmark can provide simple yet relatively robust early-stage cost estimates. However, being able to predict the required water and sewer infrastructure components before obtaining an answer that is only related to cost holds obvious benefits. This research proposes a new tool requiring limited available information, which could be applied to future development areas for estimating the likely required WDS pipeline infrastructure. For estimating the number of valves and associated infrastructure required, Liu & Kang (2021) researched typical approaches towards valve spacing and proposed an optimised approach allowing a reduction in the number of valves required without decreasing network resilience. However, applying Liu & Kang's approach to an outcome from this study, which will not necessarily include a network layout, may prove to be challenging.

Various other optimisation algorithms have been developed that are beyond the scope of this research, for example, the impact of problem formulations, pipe selection methods and optimisation algorithms on the rehabilitation of existing water distribution systems (Wang *et al.* 2020). Transient flow modelling, which has been reviewed in significant detail by Duan *et al.* (2020), also falls beyond the scope of this research. A summary of the tools developed through earlier research that are related to this research is provided in Table 1.

Scope and limitations

The focus of this study was on the potable water supply infrastructure required for future development areas, based on statistical analyses of certain physical and hydraulic parameters of existing WDSs. The study was limited to pipeline infrastructure and the occurrence of other structures (e.g. reservoirs, water towers, pumps and control valves) were not included.

Objectives

The objectives for developing the tool were to:

- Identify physical and hydraulic pipe network parameters that may influence the total pipeline length and diameter distribution of a water supply system, in terms of the known characteristics of the development area itself, which can be quantified at the early stages of a future development area project.
- Obtain a suitable sample space of existing WDSs for which all said parameters are known.
- Generate a regression model expressing the total pipeline length as a function of the other parameters.
- Generate the pipeline diameter distribution for different types of networks, for disaggregating the total pipeline length into lengths per diameter category.
- Verify and validate the model.

METHODS

This study involved applied research, where empirical methods were employed in order to solve the practical research problem. The study relied on quantitative data collection. Data were extracted from existing, calibrated hydraulics models of

Table 1 | Earlier tools for predicting required water and sewer pipeline infrastructure

Description	Reference
Table linking sewer pipe lengths and diameters to population size	Dames & Moore (1978)
Virtual infrastructure benchmarking tool to generate complex case studies for urban water systems	Sitzenfrei *et al.* (2010a, 2010b)
Case study on 30 cities to show that the relationship between pipeline length and the number of pipes by length class can be defined by the power law	Venkatesh & Brattebø (2011)
A model to predict the distribution of water pipeline lengths based on road layouts	Kobayashi *et al.* (2011)
A generic model of length, diameter distribution and replacement costs for the sewer network in a settlement with fixed area	Maurer *et al.* (2012)
A model estimating total length, total pipelines mass and diameter distribution for sewer networks where only area and population density are known	Pauliuk *et al.* (2014)
Empirical equations developed through regression analyses linking the total installation cost of sewer networks to the population size	Balaji *et al.* (2015)
A cost benchmark for water services which provides typical unit costs of water services projects and individual infrastructure components	DWS (2016)
A graph theory-based methodology for sewer system optimisation, that generates a viable sewer network layout	Turan *et al.* (2019)
A graph theory-based framework for sewer system layout and a generic scheme for decentralised layouts in both steep and flat terrains	Hesarkazzazi *et al.* (2022)
A spatial algorithm for generating simplified sewer networks which represent key characteristics of real systems, using basic topographic, demographic and urban characteristics	Duque *et al.* (2022)
A multiple linear regression tool to estimate the total sewer pipeline length for a service zone using basic service zone characteristics	Winter *et al.* (2022)

various WDSs in South Africa. The empirical evidence was subjected to statistical analysis in order to develop the model for estimating WDS pipe length as a function of basic greenfield development parameters.

The methodology encompassed several steps: (i) data collection, (ii) parameter extraction, (iii) developing and testing the model through multi-linear regression techniques and finally (iv) segregation of the total pipeline length into pipe diameter categories.

Data collection

Data collection involved data source and sample network identification and extraction, followed by selecting and abstracting the parameters of interest from each network.

Data source and sample network extraction

All the WDS models used as part of this research were at the time used in parallel by professionally registered civil engineers at GLS Consulting (www.gls.co.za) to conduct water master planning for various clients across South Africa. The hydraulic models used in this study were obtained directly from collaborators at GLS Consulting (GLS). The model nodes were already populated with water demand (node outputs and codes for land use and zoning). All typical pipeline information was available for every pipe section such a length, nominal diameter and roughness coefficient. The associated land-use information allowed further sub-sectoring of models by predominantly industrial, commercial or residential land use, for example.

From the available model data, 170 WDS models were investigated for possible inclusion in the analyses. Of these, 141 were found to be predominantly homogenous in terms of residential land use, and these were subsequently divided into two dominant land use categories, thus obtaining 90 ‘General Residential’ and 51 ‘Low-Income Residential’ models. Separate models were developed for each dominant land use category.

Parameters of interest

Table 2 lists the relevant parameters and a comment to describe how each parameter was defined. Land use was considered by classifying each model according to the predominant land use, as type A (general residential) or type B (low-income residential).

Terrain models were constructed for all 141 WDSs. The range between the highest and lowest nodal elevations would be insufficient to describe the terrain. The reason for this is that the range would not account for the number of smaller hills and terrain fluctuations inside each WDS. The range and standard deviation of the node elevations were used instead, in order to describe the terrain. By including the standard deviation, the fluctuations of the nodal elevations were accounted for. Elevation index tables, represented in Tables 3 and 4, were used to categorise each WDS. The range and standard deviation

Table 2 | Model parameters

Parameter	Unit	Definition
Total pipeline length	km	Sum of the lengths of all individual pipes in the WDS. Total pipeline length per diameter was also recorded.
Peak flow rate	L/s	Hydraulic models were populated with the hourly peak flow rate, which is derived from the average annual daily demand (AADD). The AADD is widely used for problems relating to research and design in South Africa and is also used in other Southern African countries, for example, in Malawi (Makwiza & Jacobs 2016). The minimum pressure during peak hourly demand is widely used when considering minimum system pressure (Ghorbanian *et al.* 2016).
Land area	ha	The area of each WDS was approximated by an ellipse, the major axis d_1 is the line joining the two furthest points, and the minor axis d_2 is the longest possible perpendicular bisector of d_1. The area of the ellipse is then: $\frac{\pi}{4} \times d_1 \times d_2$.
Area shape factor	–	Defined as the ratio of d_1 to d_2, describing the elongation of the ellipsoidal area.
Terrain index	–	The average value between the range and standard deviation indices for the WDS, as discussed below and determined using Tables 3 and 4.
Reservoir distance from area centroid	m	Distance between coordinates of reservoir and ellipse centroid.
Average static system pressure	m	Difference in height between reservoir full supply level and mean elevation of all WDS model nodes.

Table 3 | Elevation range index

Range index	Elevation range between highest and lowest network nodes (m)
1	10–40
2	41–70
3	71–100
4	101–130
5	131–160

Table 4 | Elevation standard deviation index

Standard deviation Index	Standard deviation of all nodal elevations (m)
1	0–6.0
2	6.1–12.0
3	12.1–18.0
4	18.1–24.0
5	24.1–30.0

index value were the same for the majority of the WDS models. For zones where this was not the case, an average value was used for terrain classification.

Final dataset

The final dataset comprised 90 'General Residential' and 51 'Low-Income Residential' data sets, for which total pipeline length, peak flow, land area, area shape, terrain index, reservoir distance from centroid and reservoir elevation above mean terrain elevation, were known.

Regression analysis

Selection of regression model

A multiple linear regression model is a linear regression model that has more than one independent variable. This typically takes the form given in Equation (1), where β_0 denotes the intercept, and β_1 to β_n denote the regression coefficients for the independent variables x_1 to x_n. Each regression coefficient represents the change in y for a unit change in the associated independent variable, if the other independent variables are kept constant (Montgomery & Runger 2014).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \quad (1)$$

The most common method for determining the intercept and regression coefficients is ordinary least squares (OLS) regression. An OLS model must satisfy five assumptions, namely, (a) lack of multi-collinearity, meaning the independent variables should be uncorrelated to each other, (b) normality, meaning the errors or residuals should be normally distributed, (c) linearity, meaning the true relationship between the dependent and independent variables should be linear in nature, (d) homoscedasticity, meaning the residuals should be independent of the values of the dependent or independent variables and (e) independence, meaning the residuals should be unrelated to their order of observation (De Veaux *et al.* 2011). When building an OLS model, a preliminary model is built using all of the candidate variables, or independent variables identified as being potentially significant. Then, the p-value for each candidate variable in the model is considered, where the p-value represents the probability that the variable is statistically insignificant (Montgomery & Runger 2014). Any variable with a p-value exceeding the selected value (0.05 in this study for a significance level of 95%) is removed, and the model is regenerated using the remaining significant variables to obtain the final model.

Selection of variables

A multiple linear regression model was developed with the total pipeline length as y, and the remaining parameters of interest from Table 2 as the candidate x-variables. The procedure was repeated for each land use. Before the preliminary model could be built, it had to be verified that multi-collinearity did not exist between the independent variables. Table 5 presents a correlation matrix for the candidate variables and indicates that the independent variables peak flow and area size are highly correlated with a correlation coefficient of 0.79. Multi-collinearity was addressed by retaining only the variable with the highest individual correlation to the total pipeline length, namely peak flow, which reduced the number of candidate independent variables to five. A preliminary regression model was then built, which would be refined to arrive at the final model. Before interpreting the performance results of any model, it was verified that the OLS assumptions were met. Linearity was indicated by the absence of curvature in partial regression plots (De Veaux *et al.* 2011) and scatter plots between the dependent and

Table 5 | Correlation matrix for candidate variables

	Total pipeline length	Peak flow	Area	Reservoir distance from centroid	Shape ratio	Reservoir height above mean	Terrain
Total pipeline length	1.00						
Peak flow	0.91	1.00					
Area	0.90	0.79	1.00				
Reservoir distance from centroid	0.37	0.33	0.36	1.00			
Shape ratio	−0.07	−0.07	−0.12	0.00	1.00		
Average static system pressure	0.14	0.11	0.19	0.58	−0.12	1.00	
Terrain	0.27	0.22	0.39	0.16	−0.06	0.43	1.00

independent variables; plots of the residuals versus each model variable also needed to display a random distribution. Independence was indicated by a random distribution when plotting the residuals versus the order of observation. Normality was indicated by the presence of a normal distribution in a histogram of the residuals, as well as the presence of a reasonably straight line on a normal probability plot. Homoscedasticity is generally indicated by the absence of any widening or narrowing in plots of the residuals versus each model variable. This final verification revealed that heteroscedasticity was present in the model, since the size of the residuals increased for datapoints with higher total pipeline length. The heteroscedasticity was addressed by introducing weighted least squares (WLS) regression, a variation of OLS, in which the larger residuals are down-weighted to reduce their disproportional impact on the regression coefficients.

Model development

For each land use category, the following process was then used to develop the final models. Scatter plots of the dependent versus each independent variable were inspected to identify and remove extreme-value points as outliers. Partial regression plots, which illustrate the relationship between the dependent and each independent variable after the effects of the other independent variables have been accounted for (De Veaux *et al.* 2011), were inspected. This inspection served to identify and remove any overly-influential points as outliers, as well as to visually assess the significance of each candidate independent variable on the independent variable. From the 'General Residential' and 'Low-Income Residential' land uses, one and six outliers were removed, respectively. From the remaining points, 20% were then randomly removed to be reserved for validity testing, leaving 80% to form the training set for model development. For each land use, a preliminary OLS model was then built, and the significant variables with $p < 0.05$ were identified to be used in the final models. The final models were built using both OLS and WLS with three different weighting systems, resulting in four final models for each land use. Subsequently, provided the five assumptions of OLS were satisfied, these models were compared in terms of the log-likelihood, AIC (Akaike's information criteria) and BIC (Bayesian information criteria) to determine the best-performing model for each land use. These likelihood indicators as well as how the results are to be interpreted are presented in Table 6.

It is noted that these likelihood-based indicators have no specific meaning and imply nothing about how good a single model is. Instead, these indicators can only be interpreted as relative values between models. Furthermore, the indicators are only applicable between models developed using the same sample points, and the same dependent variable. The selected best-performing model and performance evaluation for each land use are presented in the Results section.

Diameter distribution analysis

Analyses of the pipeline diameters and associated lengths from the 141 WDSs were used to determine typical pipeline diameter distributions. Firstly, the WDSs were classified into different area size and topography categories, as defined in Table 7. Apart from the categorisation as per Table 7, the WDSs were also classified as General Residential or Low-Income Residential. It is noted that some overlap existed between the area size and the area topography. For example, an area could be classified as both small and hilly. Moreover, a flat area could be classified as either General Residential or Low-Income Residential. The diameter distributions were obtained by determining the average diameter distribution by overall pipeline length within categories of WDSs with similar characteristics. The final distributions are presented in the Results section.

Table 6 | Indicators used for model comparison

Indicator	Interpretation
Log-Likelihood	• The log-likelihood is an alternative goodness-of-fit metric to R^2. • A higher log-likelihood implies a better model. • Like R^2, it is biased towards models with more independent variables.
Akaike's Information Criterion (AIC)	• The AIC is a goodness-of-fit metric based on the same principle as log-likelihood, but with a penalty applied for more independent variables; thus, it balances model performance and complexity. • A lower AIC implies a better model, where a two-point difference is considered significant.
Bayesian Information Criterion (BIC)	• The BIC is similar to AIC, but with a heavier penalty applied for more independent variables. • A lower BIC implies a better model; where a two-point difference is considered significant.

Table 7 | Development area category definition

Development area category	Definition
Small areas	Area <2,000 ha
Medium areas	2,000 ha < Area < 4,000 ha
Large areas	Area > 4,000 ha
Flat areas	Terrain index ≤ 2
Partially hilly areas	2 < Terrain index < 4
Hilly areas	Terrain index ≥ 4

RESULTS AND DISCUSSION

The results were generated considering 141 WDSs of at least 80 hectares in area, with a peak flow rate of less than 450 L/s, therefore can only be considered applicable to WDSs within this range.

Total pipeline length models

The only significant variable for modelling the total pipeline length was the peak flow rate, for both land uses. For the 'General Residential' land use, the reservoir distance from centroid variable exhibited a mild correlation with the dependent variable. However, this correlation was not strong enough to significantly improve the model. Therefore, for both land uses, the total pipeline length was modelled as a function of the peak flow rate. The final model formulae are presented in Equations (2) and (3), with the symbols defined in Table 8.

$$\text{General Residential:} \quad y = 4.178 + 0.317x_1 \tag{2}$$

$$\text{Low} - \text{Income Residential:} \quad y = 8.703 + 0.329x_1. \tag{3}$$

In terms of a physical interpretation, the models indicate the expected outcome that total pipeline length increases with increasing peak flow rate. The rate of increase is similar for both land use categories, in the order of 0.3 km increase in pipeline length per 1 L/s increase in peak flow rate. The 'Low-Income Residential' model has a higher intercept, indicating a denser pipeline network.

Table 9 presents the R^2 and mean absolute percentage error (MAPE) values for the training and test datasets. R^2 provides a useful and intuitive representation of the model strength, while MAPE indicates the average size of the absolute errors as percentages of the observed y-values. The test dataset values represent the model performance on data which were not used to train the model in its development. The R^2 values were considered to be good, exceeding 0.7 for both land uses in the training set, with the test values validating this strong performance when new data were considered. A visual

Table 8 | Model variables definitions

Symbol	Variable	Unit	Definition
y	Total pipeline length	km	Table 1
x_1	Peak flow rate	L/s	Table 1

Table 9 | Training and test data R^2 and MAPE for total pipeline length models

	R^2		MAPE (%)	
Land use category	Training data	Test data	Training data	Test data
General residential	0.85	0.82	31.9	32.6
Low-income residential	0.74	0.98	40.7	35.3

representation of the R^2 is presented in Figure 1 in terms of scatter plots of the predicted versus observed y-values. However, despite the strong correlation between the predicted and observed values, the MAPE values indicate that the model errors are in the order of 30% for the 'General Residential' model, and in the order of 40% for the 'Low-Income Residential'. The relative size of the errors, also visible in Figure 1, suggests that the results should be interpreted conservatively. In view of the application of this model, at the early stages of a development, many unknowns would be introduced to the planning process. The relatively large model errors are considered acceptable given the other inherent uncertainties at this early stage of a development.

Pipeline diameter distribution

The pipeline diameter distribution charts for various categories of similar WDSs are presented in Figure 2. The diameter distribution charts can be used to disaggregate the total pipeline length estimated using Equation (2) or Equation (3) into estimated length per diameter. The percentages of pipes with a diameter of ≥200 mm increase with WDS area size. The results show that small-, medium- and large areas have 9, 12 and 15% pipes with a diameter ≥200 mm, respectively. The diameter distribution categories presented below are not mutually exclusive. When an area is deemed both small and hilly, for example, the distributions within applicable categories can be averaged to obtain the most likely distribution for a particular WDS.

Application potential

The model is particularly useful during the early stages of a greenfield development when a cadastral layout may not be available yet. The average water demand of the proposed greenfield development could be estimated as soon as the population served, the number of housing units to be developed or the development footprint area is approximately known. The average water demand could be determined using typical daily demand per person, per housing unit or per land area unit. The average

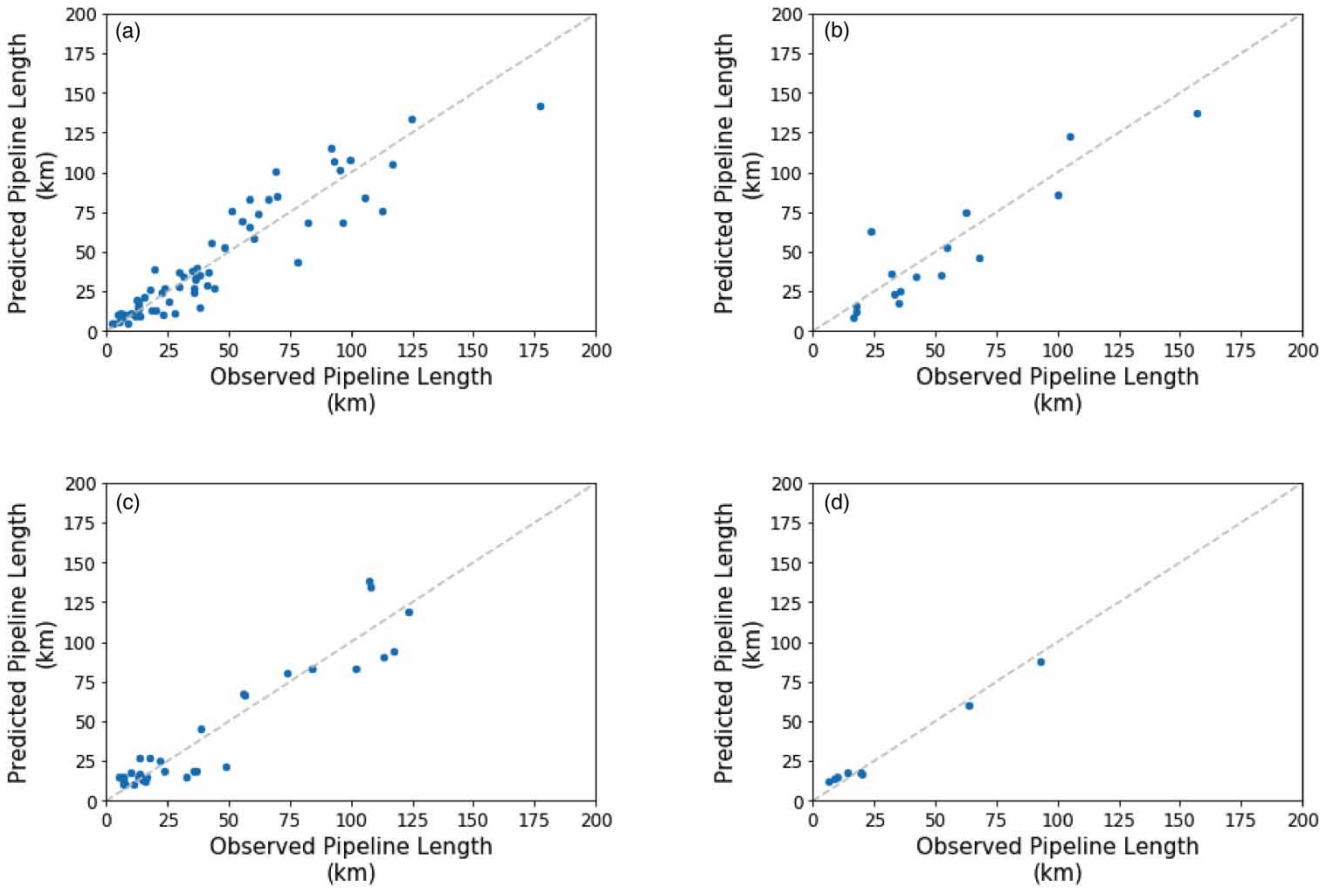


Figure 1 | Model predicted versus observed values. (a) General residential training data. (b) General residential test data. (c) Low-income residential training data. (d) Low-income residential test data.

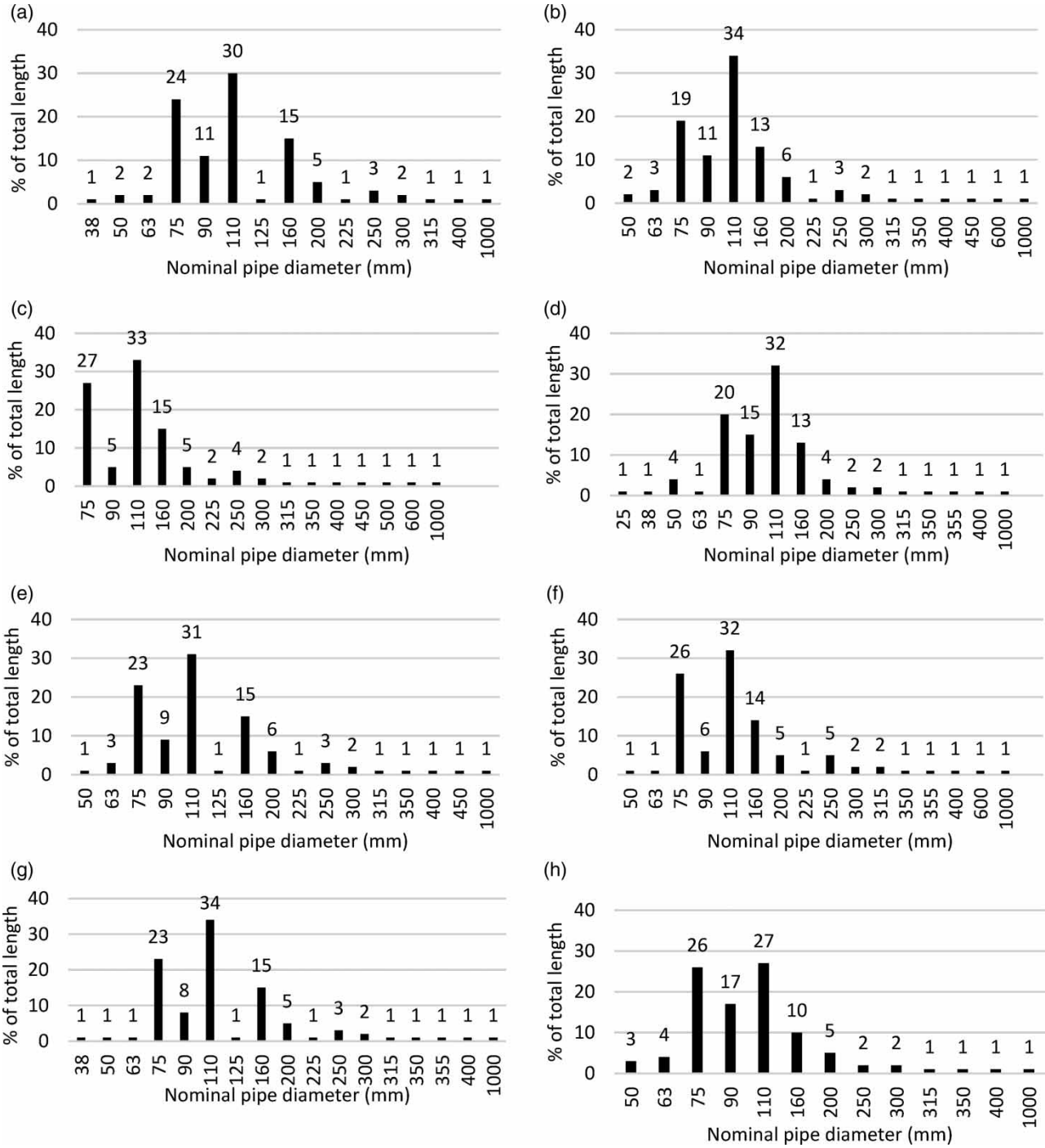


Figure 2 | Pipeline diameter distribution charts. Small areas (a), medium areas (b), large areas (c), flat areas (d), partially hilly areas (e), hilly areas (f), general residential areas (g), low-income residential areas (h).

water demand thus determined could be multiplied by a peak hour factor associated with this type of development in order to estimate the peak hour flow rate.

The model enables the prediction of the total length of WDS pipes to service a future development area as a function of the peak flow rate. Once the total pipeline length is determined, the diameter distribution charts can be used to disaggregate the total pipeline length into estimated length per diameter. With the approximate total length of pipework and associated length per diameter known, a budget estimate is enabled by the utility for the future construction of planned assets.

Loubser *et al.* (2021) highlighted the widespread occurrence of intermittent water supply (IWS) in South Africa. One of the causes of IWS is insufficient WDS pipes, in view of the population served by the particular WDS (Maake & Holtzhausen 2015). The model developed during this research can potentially be used to analyse systems subjected to IWS. The model could be employed, for example, to compare the expected WDS pipe length (as function of the estimated total peak hour flow rate) to the actual WDS pipe length of the system subjected to IWS. Systems subjected to IWS could be expected to have insufficient pipe length when compared to the model results or display diameter distributions that vary drastically from the model outcomes (for example, when relatively higher percentages of smaller diameter pipelines are present). Therefore, the model would allow a user to ascertain whether an existing water supply network has potentially been stretched beyond its design capacity. A similar result can be achieved via detailed hydraulic modelling of the WDS, yet this tool provides an indication of capacity problems with only a few inputs – a task that is made possible with limited resources and subject to stringent budget- and time constraints.

CONCLUSIONS

A model enabling the prediction of the water supply pipeline infrastructure required for future development areas was successfully developed. It was found that peak hour flow rate dominated the result in terms of total WDS pipe length. With the peak hour demand of an existing or future residential development area estimated or known, the model can be applied to crudely estimate the total WDS pipe length required. Subsequently, the diameter distribution charts can be used to disaggregate the total pipeline length into estimated length per diameter.

The model could find application in estimating the water supply pipeline infrastructure required for greenfield developments at an early stage. Moreover, the model can be used to estimate the maximum capacity of existing water networks, in order to ascertain whether the WDS can sustain certain maximum peak hour demands – failing which would require infrastructure upgrades, or else could result in low pressures and ultimately IWS.

Multiple linear regression was employed to develop a model based on South African residential water supply network data. Thus, care should be exercised when using the model outside the South African context and on non-residential networks. The model is applicable to predominantly residential service zones with a footprint $\geq$80 hectares and a peak hour flow rate of $\leq$450 L/s.

Future research should focus on modelling smaller areas and water distribution zones, which may require unique models associated with area size categories $<$80 hectares. This would enable wider practical application of the model in smaller developments. In addition, given the good correlation between development area size and total pipeline length, a second model selecting area size as the dominant independent parameter could also be developed. In order to generalise the model for a wider geographic application, the model could be extended to include large datasets of water distribution systems from other parts of the world.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Balaji, B., Mariappan, P. & Senthamilkumar, S. 2015 A cost estimate model for sewerage system. *ARPN Journal of Engineering and Applied Sciences* **10**, 1–6.

Dames & Moore 1978 *Construction Costs for Municipal Wastewater Conveyance Systems: 1973–1977*, US EPA Technical Report, Office of Water, EPA 430/9-77-014.

De Veaux, R., Velleman, P. & Bock, D. 2011 Chapter 30 Multiple regression. In: *Stats: Data and Models*, 3rd edn, Pearson, Boston, pp. 30.1–30.23.

Duan, H.-F., Pan, B., Wang, M., Chen, L., Zheng, F. & Zhang, Y. 2020 State-of-the-art review on transient flow modeling and utilization for urban water supply system (UWSS). *Journal of Water Supply: Research and Technology – AQUA* **69**, 8. doi:10.2166/aqua.2020.048.

Duque, N., Bach, P. M., Scholten, L., Fappiano, F. & Maurer, M. 2022 A simplified sanitary sewer system generator for exploratory modelling at city-scale. *Water Research* **209**. doi:10.1016/j.watres.2021.117903.

DWS 2016 *Cost Benchmark for Water Services Projects*. Department of Water and Sanitation, Pretoria.

Ghorbanian, V., Karney, B. & Guo, Y. 2016 Pressure standards in water distribution systems: reflection on current practice with consideration of some unresolved issues. *Journal of Water Resources Planning and Management* **142**, 04016023.

Hesarkazzazi, S., Hajibabaei, M., Bakhshipour, A. E., Dittmer, U., Haghighi, A. & Sitzenfrei, R. 2022 Generation of optimal (de)centralised layouts for urban drainage systems: a graph-theory-based combinatorial multi-objective optimization framework. *Sustainable Cities and Society* **81**. doi:10.1016/j.scs.2022.103827.

Kobayashi, T., Yamazaki, F. & Nagata, S. 2011 Estimation of the distribution of water pipeline's length for earthquake damage assessment based on other infrastructure data. *Journal of Social Safety Science* **15**, 163–168.

Liu, J. & Kang, Y. 2021 Segment-based resilience response and intervention evaluation of water distribution systems. *AQUA – Water Infrastructure, Ecosystems and Society* **71** (1). doi:10.2166/aqua.2021.133.

Loubser, C., Mwiinga Chimbanga, B. & Jacobs, H. E. 2021 Intermittent water supply: a South African perspective. *Water SA* **47** (1). doi:10.17159/wsa/2021.v47.i1.9440.

Maake, M. T. & Holtzhausen, N. 2015 Factors affecting the provision of sustainable water services in the Mopani District Municipality, Limpopo Province. *Administratio Publica* **23** (4),), 248–271.

Makwiza, C. & Jacobs, H. E. 2016 Assessing the impact of property size on the residential water use for selected neighbourhoods in Lilongwe, Malawi. *Journal Water Sanitation and Hygiene for Development* **06**, 242–251. doi:10.2166/washdev.2016.014.

Maurer, M., Scheidegger, A. & Herlyn, A. 2012 Quantifying costs and lengths of urban drainage systems with a simple static sewer infrastructure model. *Urban Water Journal* **10** (4), 268–280.

Moeini, R. & Afshar, M. H. 2018 Hybridizing ant colony optimization algorithm with nonlinear programming method for effective optimal design of sewer networks. *Water Environment Research* **91** (4), 300–321.

Montgomery, D. C. & Runger, G. C. 2014 *Applied Statistics and Probability for Engineers*, 6th edn. John Wiley & Sons, Singapore.

Pauliuk, S., Venkatesh, G., Brattebø, H. & Müller, D. B. 2014 Exploring urban mines: pipe length and material stocks in urban water and wastewater networks. *Urban Water Journal* **11** (4), 274–283.

Sitzenfrei, R., Fach, S., Kinzel, H., Urich, C. & Rauch, W. 2010a A multilayer cellular automata approach for algorithmic generation of virtual case studies – VIBe. *Water Science & Technology* **61** (1), 37–45.

Sitzenfrei, R., Fach, S., Kleidorfer, M., Urich, C. & Rauch, W. 2010b Dynamic virtual infrastructure benchmarking: DynaVIBe. *Water Science & Technology: Water Supply* **10** (4), 600–609.

Turan, M. E., Bacak-Turan, G., Cetin, T. & Aslan, E. 2019 Feasible sanitary sewer network generation using graph theory. *Advances in Civil Engineering* **2019**. doi:10.1155/2019/8527180.

Urich, C., Sitzenfrei, R., Möderl, M. & Rauch, W. 2010 An agent-based approach for generating virtual sewer systems. *Water Science & Technology* **62** (5), 1090–1097.

Venkatesh, G. & Brattebø, H. 2011 Testing the power law on urban water and wastewater pipeline networks. *Vatten* **67**, 157–160.

Wang, Q., Huang, W., Yang, X., Wang, L., Wang, Z. & Wang, Y. 2020 Impact of problem formulations, pipe selection methods and optimization algorithms on rehabilitation of water distribution systems. *Journal of Water Supply: Research and Technology – AQUA*. **69**, 8. doi:10.2166/aqua.2020.053.

Winter, J. M., Loubser, C. & Bosman, A. 2022 Estimating sanitary sewer pipeline infrastructure from basic characteristics of a service zone. *Water SA* **48** (2), 161–170. doi:10.17159/wsa/2022.v48.i2.3900.

First received 7 November 2022; accepted in revised form 8 December 2022. Available online 20 December 2022