Zecchin, A., Lambert, M., Simpson, A.R., White, L.

Parameter identification in pipeline networks: transient-based expectation-maximization approach for systems containing unknown boundary conditions

Journal of Hydraulic Engineering, 2014; 140(6):04014020-1-04014020-12

http://dx.doi.org/10.1061/(ASCE)HY.1943-7900.0000849

Parameter identification in pipeline networks: transient-based expectation-maximisation approach for systems containing unknown boundary conditions

by

Zecchin, A.C, Lambert, M.F., Simpson, A.R. and White, L.

Journal of Hydraulic Engineering

Citation:
Zecchin, A.C, Lambert, M.F., Simpson, A.R. and White, L. (2014). "Parameter identification in pipeline networks: transient-based expectation-maximisation approach for systems containing unknown boundary conditions". *Journal of Hydraulic Engineering*, ASCE, Feb., Vol. 140, No. 6, 04014020, DOI: 10.1061 /(ASCE) HY. 1943-7900.0000849.

For further information about this paper please email Angus Simpson at angus.simpson@adelaide.edu.au

PARAMETER IDENTIFICATION IN PIPELINE NETWORKS: A TRANSIENT BASED EXPECTATION-MAXIMISATION APPROACH FOR SYSTEMS CONTAINING UNKNOWN BOUNDARY CONDITIONS

A. C. Zecchin[1], M. F. Lambert[2], A. R. Simpson[3], L. B. White[4],

ABSTRACT

The simulation of hydraulic transients within fluid line networks is important for many applications (for example, water hammer analysis within distribution networks). However, in many instances, modelling efforts are impeded by the fact that the pipeline parameters are either unknown, or can vary significantly from their assumed design values. Consequently, research efforts have focused on the development of parameter identification techniques, mapping from measured transient data to pipeline parameter estimates. A limitation with previous works has been the need for systems to have all boundary conditions either measured or known (e.g. transient pressure measurements or reservoir boundary conditions). This paper aims to relax this requirement, and presents a parameter identification method for fluid line networks based on transient-state measurements of the hydraulic state variables of pressure and flow, in the presence of unmeasured and unknown boundary conditions. Utilising a Laplace-domain admittance matrix representation of the system, the contribution to the hydraulic system dynamics from the measured and unmeasured state variables (i.e. boundary conditions) is made explicit. This model is then used as the basis for the development of a parameter estimation methodology based on the expectation-maximization (EM) algo-

[1]School of Civil, Environmental and Mining Engineering, The University of Adelaide, Australia. E-mail: azecchin@civeng.adelaide.edu.au

[2]School of Civil, Environmental and Mining Engineering, The University of Adelaide, Australia.

[3]School of Civil, Environmental and Mining Engineering, The University of Adelaide, Australia.

[4]School of Electrical and Electronic Engineering, The University of Adelaide, Australia.

rithm. The importance of the EM approach is that it provides a framework for parameter estimation in the presence of unmeasured state variables, by effectively integrating out the influence of the unmeasured variables. Numerical examples demonstrate the utility of this method for a network with a range of pipeline models.

Keywords: fluid transients; pipeline networks; parameter estimation; expectation-maximization.

INTRODUCTION

The pipeline parameters of a pipe network can vary significantly from their assumed design values due to aging (*e.g.* corrosion of pipe wall material, or buildup of solids within the pipeline), imperfections in installation (*e.g.* supports not completely restraining the pipeline), and issues in manufacturing (*e.g.* variation pipeline roughness heights). The need for accurate simulation of pipeline systems, combined with the outlined parametric uncertainty, has led to significant research efforts on pipeline parameter identification methods (*e.g.* Isermann (1984), Liggett and Chen (1994), Nash and Karney (1999)), with a particular focus on leak detection (*e.g.* Liou and Tian (1995), Lee et al. (2005)).

Many of these methods have focused on approaches tailored for single pipeline systems with either measured or known boundary conditions (*e.g.* see Verde et al. (2007), Wang et al. (2002), respectively). Few methods have dealt with fluid line networks of a general topology, namely, the time-domain inverse transient method (Liggett and Chen 1994), the least squares calibration approach based on the frequency-domain impedance matrix method (Kim 2008), and the maximum likelihood estimation approach (Zecchin et al. 2013) based on the Laplace-domain network admittance matrix model (Zecchin et al. 2009; Zecchin et al. 2010). The complexities of parameter estimation within a pipeline network are associated with the large number of parameters, and the difficulty in developing an estimation algorithms to correctly use the measured transient data and the, sometimes uncertain, boundary condition information. As such, to date, a limitation with previous works has been the need for networks to have all boundary conditions either measured (through transient pressure, or flow, measurements) or known (*e.g.* a reservoir, junction or valve boundary condition). An example of such situations of uncertain boundary conditions are when the pipe network of interest is connected to a broader network through unmonitored connection points (allowing for the transient dynamics of the broader network to influence the dynamic behaviour of the network of interest), or when the network of interest is connected to hydraulic components whose properties are unknown.

This paper proposes the use of the expectation-maximisation (EM) algorithm to provide a rigorous way of dealing with the case of pipe network parameter identification in the presence of uncertain boundary conditions. The EM algorithm (Dempster et al. 1977), is a general statistical parameter estimation method used in situations where the data is incomplete, or there exist hidden state variables upon which the system dynamics depend (Michiko and Kazunori 2004).

The proposed approach utilises the Laplace-domain network admittance matrix model of Zecchin et al. (2009) to develop the identification method in the frequency-domain. The primary advantage of the frequency-domain approach is that it enables an analytic representation of the influence of the measured and unmeasured nodal states on the network dynamics, which is a critical first step for the application of the EM algorithm. Additional advantages of frequency-domain methods are that they are very computationally efficient, and they do not suffer from the grid generation difficulties associated with parameter estimation using time-domain methods (Kim 2008) (*i.e.* for the inverse transient method, the computational grid of the inverse model is dependent on the pipeline wave speed parameter, which is itself an unknown parameter requiring estimation).

As frequency-domain methods deal with linear dynamics, they provide only an approximation to the true nonlinear network behaviour, where the accuracy of the approximation is dependent on the magnitude of the flow perturbation about the steady-state (or the set-point about which the linear approximation is made) (Wylie and Streeter 1993). Despite this limitation, typically only small flow perturbations are required to an achieve adequate excitation in the pressure response of the system, meaning that frequency-domain methods have been successfully utilised for both single line (*e.g.* Lee et al. (2005), Mohapatra et al. (2006)) and network applications (*e.g.* (Zecchin et al. 2012)).

PROBLEM FORMULATION

To explain the objective of the paper, an example is first given, afterwhich the system of network equations that govern the hydraulic state variables is outlined, and the parameter estimation problem is formally explained.

Example. *Consider the network depicted in Figure 1(a) with 13 pipes, eight junctions, a surge vessel (capacitive element), an emitter, and a reservoir. Say that the prior information for this system only describes the topology of the subnetwork comprising the first seven nodes, and 11 links of the network as depicted in Figure 1(b). The prior information indicates that there is an additional connection at node 7, but the structure of the network beyond this node is unknown (that is, as depicted in Figure 1(b), the nodal pressure and flow for this node is unknown). Each pipe within the known 11-pipe network of Figure 1(b) has a set of unknown parameters that require estimation (*e.g. *roughness, diameter, wavespeed), symbolised by the sets* $\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_{11}$. *Consider that the network is excited into a transient state by a measured flow perturbation* $\theta_4(t)$ *at node 4 (denoted* $\theta_d(t)$ *in Figure 1), and the pressure response of the network is measured at nodes 2, 3, 4 and 6. The parameter identification consists of estimating the pipeline parameter values* $\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_{11}$ *given the pressure measurements* $\psi_2(t)$, $\psi_3(t)$, $\psi_4(t)$, $\psi_6(t)$ *at nodes 2, 3, 4 and 6, and* $\theta_4(t)$ *at* $t = \Delta t, \ldots N\Delta t$, *and the known boundary conditions of the pipeline interactions at the six junctions and the reservoir.*

What separates this work from previous pipe network parameter estimation methodologies (Kim 2008; Zecchin et al. 2013) is the presence of boundary nodes within the network for which neither the pressure or flow are measured, creating an uncertain boundary condition (*e.g.* node 7 in this case).

Network Equations

Following the notation of Zecchin et al. (2013), it is convenient to describe a network as a connected graph $\mathcal{G}(\mathcal{N}, \Lambda)$ (Diestel 2000) consisting of the node set $\mathcal{N} = \{1, 2, ..., n_n\}$, and the link set $\Lambda = \{\lambda_1, \lambda_2, ..., \lambda_{n_\Lambda}\}$ ($\lambda_j = (i,k)$ where i and k are the upstream and downstream nodes of link j). Each node is associated with a junction that is connected to a number of links, and each link is associated with a distributed pipe element where the directed nature of the link describes the positive flow direction sign convention of the element. With the given notation, a fluid line network is defined as the pair $(\mathcal{G}(\mathcal{N},\Lambda),\mathcal{P})$ where $\mathcal{G}(\mathcal{N},\Lambda)$ is the network graph of nodes $\mathcal{N}$ and links Λ, and $\mathcal{P} = \{\mathcal{P}_j : \lambda_j \in \Lambda\}$ is the set of pipeline coefficients and operators that describe the dynamics of each pipe j.

The state space of the network $(\mathcal{G}(\mathcal{N},\Lambda),\mathcal{P})$ is given by the distributions of pressure and flow along each line of the network, and the imposed nodal states of pressure and flow. The distributed line states are given by

$$\boldsymbol{p}(\boldsymbol{x},t) = \begin{bmatrix} p_1(x_1,t) \\ \vdots \\ p_{n_\Lambda}(x_{n_\Lambda},t) \end{bmatrix}, \quad \boldsymbol{q}(\boldsymbol{x},t) = \begin{bmatrix} q_1(x_1,t) \\ \vdots \\ q_{n_\Lambda}(x_{n_\Lambda},t) \end{bmatrix}, \tag{1}$$

respectively, where $\boldsymbol{x} = [x_1 \cdots x_{n_\Lambda}]^T$ is the vector of spatial coordinates, (*i.e.* $\boldsymbol{x} \in \mathcal{X} = \mathcal{X}_1 \times \cdots \mathcal{X}_{n_\Lambda}$ where $\mathcal{X}_j = [0, l_j]$), $t \in \mathbb{R}$ is time, and n_Λ is the number of links. The nodal states (imposed by the connected pipelines) are given by

$$\boldsymbol{\psi}(t) = [\psi_1(t) \cdots \psi_{n_n}(t)]^T, \quad \boldsymbol{\theta}(t) = [\theta_1(t) \cdots \theta_{n_n}(t)]^T. \tag{2}$$

where ψ_i and θ_i are the pressure at node i and the flow into node i respectively (for reasons of passivity, the flow sign convention is taken as positive when directed into the network). As outlined in Zecchin et al. (2009), the states (1) and (2) are governed by the following system of equations

$$\frac{\partial \boldsymbol{p}}{\partial \boldsymbol{x}} + \boldsymbol{R}_o \left(\frac{\partial \boldsymbol{q}}{\partial t} + \boldsymbol{\mathcal{R}}\left(\boldsymbol{q}\right) \right) = \mathbf{0}, \quad \boldsymbol{x} \in \boldsymbol{\mathcal{X}} \tag{3}$$

$$\frac{\partial \boldsymbol{q}}{\partial \boldsymbol{x}} + \boldsymbol{C}_o \left(\frac{\partial \boldsymbol{p}}{\partial t} + \boldsymbol{\mathcal{C}}\left(\boldsymbol{p}\right) \right) = \mathbf{0}, \quad \boldsymbol{x} \in \boldsymbol{\mathcal{X}} \tag{4}$$

$$\begin{pmatrix} \boldsymbol{N}_u & -\boldsymbol{N}_d \end{pmatrix} \begin{pmatrix} \boldsymbol{q}(\boldsymbol{x} = \mathbf{0}) \\ \boldsymbol{q}(\boldsymbol{x} = \boldsymbol{l}) \end{pmatrix} = \boldsymbol{\theta} \tag{5}$$

$$\begin{pmatrix} \boldsymbol{p}(\boldsymbol{x} = \mathbf{0}) \\ \boldsymbol{p}(\boldsymbol{x} = \boldsymbol{l}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{N}_u & \boldsymbol{N}_d \end{pmatrix}^T \boldsymbol{\psi} \tag{6}$$

where $\boldsymbol{R}_o = \operatorname{diag}\left[R_1 \cdots R_{n_\Lambda}\right]$ is the matrix of resistance coefficients, $\boldsymbol{\mathcal{R}}(\boldsymbol{q}) = \operatorname{diag}\left[\mathcal{R}_1(q_1) \cdots \mathcal{R}_{n_\Lambda}(q_{n_\Lambda})\right]$ is the matrix of resistance operators, $\boldsymbol{C}_o = \operatorname{diag}\left[C_1 \cdots C_{n_\Lambda}\right]$ is the matrix of compliance coefficients, $\boldsymbol{\mathcal{C}}(\boldsymbol{p}) = \operatorname{diag}\left[\mathcal{C}_1(p_1) \cdots \mathcal{C}_{n_\Lambda}(p_{n_\Lambda})\right]$ is the matrix of compliance operators, and $\boldsymbol{N}_u$ and $\boldsymbol{N}_d$ are upstream and downstream incidence matrices that describe the connectivity of the network, and are given by

$$\left\{\boldsymbol{N_u}\right\}_{i,j} = \begin{cases} 1 & \text{if} \quad \lambda_j \in \Lambda_{u,i} \\ 0 & \text{otherwise} \end{cases}, \qquad \left\{\boldsymbol{N_d}\right\}_{i,j} = \begin{cases} 1 & \text{if} \quad \lambda_j \in \Lambda_{d,i} \\ 0 & \text{otherwise} \end{cases},$$

where the sets $\Lambda_{ui} = \{(i,k), k \in \mathcal{N} : (i,k) \in \Lambda\}$ and $\Lambda_{di} = \{(k,i), k \in \mathcal{N} : (k,i) \in \Lambda\}$ correspond to the set of links directed from and to node i respectively.

Equations (3) and (4) describe the mass and momentum conservation for the flow within a fluid line, respectively, and (5) and (6) describe the interaction between the link end points and the nodal states through the nodal mass and nodal pressure conservation equations, respectively (note that $\boldsymbol{x} = \mathbf{0}$ symbolises the upstream point in each of the lines, and $\boldsymbol{x} = \boldsymbol{l}$ symbolises the downstream point).

The parametric dependency of the dynamics for pipeline j are characterized through the pipeline coefficients and operators $\mathcal{P}_j = \{(R_j, \mathcal{R}_j), (C_j, \mathcal{C}_j)\}$. These are dependent on the parameter set $\boldsymbol{\vartheta}_j$ which includes the physical parameters such as pipe diameter, length, wavespeed, and roughness (that is $R_j = R_j(\boldsymbol{\vartheta}_j)$, $\mathcal{R}_j = \mathcal{R}_j(\boldsymbol{\vartheta}_j)$, $C_j = C_j(\boldsymbol{\vartheta}_j)$, and $\mathcal{C}_j = \mathcal{C}_j(\boldsymbol{\vartheta}_j)$). The operators $\mathcal{R}_j$ and $\mathcal{C}_j$ are typically integrodifferential (and possibly non-linear) and many different forms exist based on different assumptions about the underlying partial differential equation (PDE) system (Rieutord and Blanchard 1979; Stecki and Davis 1986; Vardy and Brown 2007). Two different forms for $\mathcal{R}$ are used within the numerical experiments outlined later.

Definition of Parameter Identification Problem

The entire parameter set requiring estimation for the network $(\mathcal{G}(\mathcal{N},\Lambda),\mathcal{P})$ is given by the set $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}_1 \cup \cdots \cup \boldsymbol{\vartheta}_{n_\Lambda}$. The parameter identification problem can be formally outlined as follows. Given a network $(\mathcal{G}(\mathcal{N},\Lambda),\mathcal{P})$ with unknown parameter values $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}^*$, the network parameter identification problem is defined as identifying the most likely parameter estimate $\widehat{\boldsymbol{\vartheta}}$ within the parameter space $\boldsymbol{\Upsilon}$, from the measurement set

$$\left\{\tilde{\boldsymbol{\psi}}_m(t), \tilde{\boldsymbol{\theta}}_m(t) : t = 0, \Delta t, \ldots, N\Delta t\right\} \tag{7}$$

where each measurement is given by

$$\begin{pmatrix} \tilde{\boldsymbol{\psi}}_m(t) \\ \tilde{\boldsymbol{\theta}}_m(t) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\psi}_m(t) \\ \boldsymbol{\theta}_m(t) \end{pmatrix} + \begin{pmatrix} \boldsymbol{e}_\psi(t) \\ \boldsymbol{e}_\theta(t) \end{pmatrix} \tag{8}$$

where $\boldsymbol{e}_\psi$ and $\boldsymbol{e}_\theta$ are measurement error terms, and $\boldsymbol{\psi}_m$ and $\boldsymbol{\theta}_m$ are the true values of the measured states related to the state vectors $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$ by

$$\boldsymbol{\psi}_m = \boldsymbol{A}_\psi \boldsymbol{\psi}, \quad \boldsymbol{\theta}_m = \boldsymbol{A}_\theta \boldsymbol{\theta} \tag{9}$$

where $\boldsymbol{A}_\psi$ and $\boldsymbol{A}_\theta$ are binary matrices (that pick out the relevant measured nodes from the state vectors), and $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$ are governed by the system (3)-(6), where $\boldsymbol{R}_o$, $\boldsymbol{\mathcal{R}}$, $\boldsymbol{C}_o$, and $\boldsymbol{\mathcal{C}}$

are dependent on the unknown parameter value $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}^*$.

The parameter space $\boldsymbol{\Upsilon}$ represents the *a priori* knowledge of the parameter values and is taken as bounded intervals on the real line (as the pipeline parameters are typically known to lie between upper and lower bounds), and the error terms $\boldsymbol{e}_\psi(t)$ and $\boldsymbol{e}_\theta(t)$ for $t = 0, \Delta t, \ldots$ are stationary processes (i.e. the error statistics do not change with time) with power spectra $\boldsymbol{S}_\psi(\omega)$ and $\boldsymbol{S}_\theta(\omega)$, respectively.

As outlined in the sections below, the uncertain boundary conditions serve to complicate this process as the system from which the measurements are taken possess unmeasured and unaccounted for dynamic inputs. This is further outlined in the following sections.

NETWORK REPRESENTATION

A physical model must be adopted in order to map the measurements $\tilde{\boldsymbol{\psi}}_m$ and $\tilde{\boldsymbol{\theta}}_m$ to an estimate of the parameter set $\boldsymbol{\vartheta}$, through the minimisation of an error function that indicates the *goodness of fit* between the model and the measurements. Typical approaches have adopted a least squares fitting of numerical models of (3)-(6) directly [*e.g.* the method of characteristics model adopted in the inverse transient method (Liggett and Chen 1994)], or the use of transformed linearised approximations of (3)-(6) for either a least squares fit (Kim 2008) or a maximum likelihood estimation (Zecchin et al. 2013).

Implicit in all of these approaches is that each node within the network possesses either a known boundary condition (either nodal pressure for a reservoir, or nodal flow as for a junction where $\theta = 0$), or that either one of these nodal states is measured. That is, prior to the work presented within this paper, no methods have previously been formulated to deal with the situation where there are nodes within the network for which the transient behaviour of both the nodal pressure or flow is unknown.

This section is structured as follows. First, a framework to systematically categorize the nodes, based on the information that is available from them, is outlined. Second, a Laplace-domain model is developed that decomposes the system dynamics into a term dependent on all measured nodal states, and a term dependent on all unmeasured node states. This model serves as the basis for the expectation-maximization algorithm derived in the following section.

Network nodal partitioning

For a given network $(\mathcal{G}(\mathcal{N}, \Lambda), \mathcal{P})$ containing known boundary conditions and measured nodal states, the nodes $\mathcal{N}$ can be categorized into disjoint node sets depending on whether the nodal variables are known, measured or unknown. Three disjoint subsets exist, namely

1. $\mathcal{A}$, the set of nodes for which neither of the variables of pressure and flow are known,
2. $\mathcal{B}$, the set of nodes for which the nodal flow is known, and
3. $\mathcal{C}$, the set of nodes for which the nodal pressure is known.

This partitioning can be further refined by considering combinations for which the nodal states are either measured or unmeasured. This results in the 8 unique sets that are tabulated in Table 1. Note that the following relations hold, $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2 \cup \mathcal{A}_3 \cup \mathcal{A}_4$, $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$, and $\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2$. These 8 sets represent a complete partitioning covering all realistic combinations for which the nodal variables are simultaneously known or unknown. The only omission is the case of known pressure and known flow, which is an unrealistic case as only one of these variables can be controlled and hence known (*i.e.* at a junction the outflow can be controlled, and hence it is known to be zero, but the pressure must be measured, and at a reservoir, the pressure can be controlled and is known, but the outflow must be measured).

Given the sets in Table 1, the network nodal state space can be partitioned as

$$\boldsymbol{\psi} = \begin{pmatrix} \boldsymbol{\psi}_{\mathcal{A}_1} \\ \vdots \\ \boldsymbol{\psi}_{\mathcal{A}_4} \\ \boldsymbol{\psi}_{\mathcal{B}_1} \\ \boldsymbol{\psi}_{\mathcal{B}_2} \\ \boldsymbol{\psi}_{\mathcal{C}_1} \\ \boldsymbol{\psi}_{\mathcal{C}_2} \end{pmatrix}, \qquad \boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta}_{\mathcal{A}_1} \\ \vdots \\ \boldsymbol{\theta}_{\mathcal{A}_4} \\ \boldsymbol{\theta}_{\mathcal{B}_1} \\ \boldsymbol{\theta}_{\mathcal{B}_2} \\ \boldsymbol{\theta}_{\mathcal{C}_1} \\ \boldsymbol{\theta}_{\mathcal{C}_2} \end{pmatrix} \tag{10}$$

where each of the $\boldsymbol{\psi}_X$ and $\boldsymbol{\theta}_X$ are $n_X \times 1$ vectors ($X = \mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \mathcal{A}_4, \mathcal{B}_1, \mathcal{B}_2, \mathcal{C}_1$ and $\mathcal{C}_2$). Given this partitioning, the known, measured and unmeasured variables are

$$\boldsymbol{\psi}_k = \begin{pmatrix} \boldsymbol{\psi}_{\mathcal{C}_1} \\ \boldsymbol{\psi}_{\mathcal{C}_2} \end{pmatrix}, \quad \boldsymbol{\psi}_m = \begin{pmatrix} \boldsymbol{\psi}_{\mathcal{A}_1} \\ \boldsymbol{\psi}_{\mathcal{A}_2} \\ \boldsymbol{\psi}_{\mathcal{B}_1} \end{pmatrix}, \quad \boldsymbol{\psi}_u = \begin{pmatrix} \boldsymbol{\psi}_{\mathcal{A}_3} \\ \boldsymbol{\psi}_{\mathcal{A}_4} \\ \boldsymbol{\psi}_{\mathcal{B}_2} \end{pmatrix} \tag{11}$$

for pressure, respectively, and

$$\boldsymbol{\theta}_k = \begin{pmatrix} \boldsymbol{\theta}_{\mathcal{B}_1} \\ \boldsymbol{\theta}_{\mathcal{B}_2} \end{pmatrix}, \quad \boldsymbol{\theta}_m = \begin{pmatrix} \boldsymbol{\theta}_{\mathcal{A}_1} \\ \boldsymbol{\theta}_{\mathcal{A}_3} \\ \boldsymbol{\theta}_{\mathcal{C}_1} \end{pmatrix}, \quad \boldsymbol{\theta}_u = \begin{pmatrix} \boldsymbol{\theta}_{\mathcal{A}_2} \\ \boldsymbol{\theta}_{\mathcal{A}_4} \\ \boldsymbol{\theta}_{\mathcal{C}_2} \end{pmatrix} \tag{12}$$

for the nodal flows, respectively.

Given this partitioning of the nodal set based on the information available at each node, the differentiation of this work from that of previous studies can be outlined more precisely. Namely, all previous works have dealt with networks for which neither the nodal pressure ψ nor nodal flow θ were unknown or unmeasured, that is, for all previous work $\mathcal{A}_4 = \emptyset$. The consideration of cases where $\mathcal{A}_4 \neq \emptyset$ is the primary novel contribution of this work.

Example. *Reconsider the example network in Figure 1(b) with pressure measurements at nodes 2, 3, 4 and 6, and a flow measurement at node 4. Node 1 is a reservoir, and so it has a known head but with unknown (and unmeasured) nodal flow, so* $1 \in \mathcal{C}_2$. *Nodes 2, 3, 5 and 6 are junctions, with known zero nodal flow, hence they are all in the set* $\mathcal{B}$. *Of these nodes,* $2, 3, 6 \in \mathcal{B}_1$ *as they contain pressure measurements, and* $5 \in \mathcal{B}_2$ *as the pressure is not measured at this node. Node* 4 *is in set* $\mathcal{A}_1$ *as both the nodal flow and pressure are measured at this node. Node* 7 *is in the set* $\mathcal{A}_4$ *as both the pressure and nodal flow are unknown and unmeasured at this node. The entire nodal categorisation is summarised as Scenario 1 in Table 2. As a result of this categorisation, the measured and unmeasured variables from* (11) *and* (12) *are summarised as follows:* $\boldsymbol{\psi}_k = \psi_1$; $\boldsymbol{\psi}_m = [\psi_4 \, \psi_2 \, \psi_3 \, \psi_6]^T$; $\boldsymbol{\psi}_u = [\psi_7 \, \psi_5]^T$; $\boldsymbol{\theta}_k = [\theta_2 \, \theta_3 \, \theta_6 \, \theta_5]^T$; $\boldsymbol{\theta}_m = \theta_4$; $\boldsymbol{\theta}_u = [\theta_7 \, \theta_1]^T$.

Laplace-Domain Network Admittance Matrix

For a linear network with homogeneous initial conditions (or a nonlinear network, linearized about an initial steady-state operating point), Zecchin et al. (2009) demonstrated that the nodal pressures could be mapped to the nodal flows through the following admittance map

$$\boldsymbol{\theta}(t) = \int_0^t \boldsymbol{\mathcal{Y}}(t-\tau)\boldsymbol{\psi}(\tau)\mathrm{d}\tau \tag{13}$$

where $\boldsymbol{\mathcal{Y}}$ is the network admittance matrix whose (i,k) element $\mathcal{Y}_{i,k}$ is the impulse response function for the contribution of the pressure at node k to the flow at node i. No closed form expression for $\boldsymbol{\mathcal{Y}}$ exists, but the Laplace transform of (13) is

$$\boldsymbol{\Theta}(s) = \boldsymbol{Y}(s)\boldsymbol{\Psi}(s) \tag{14}$$

where s is the Laplace variable, the uppercase symbols represent the Laplace transforms of their lower case counter parts, and for which the elemental transfer functions $Y_{i,k}$ are given by

$$Y_{i,k}(s) = \begin{cases} \sum_{\lambda_j \in \Lambda_i} Z_{cj}^{-1}(s) \coth \Gamma_j(s) & \text{if } k = i \\ -Z_{cj}^{-1}(s) \operatorname{csch} \Gamma_j(s) & \text{if } \lambda_j \in \Lambda_i \cap \Lambda_k \\ 0 & \text{otherwise} \end{cases} . \tag{15}$$

where Γ_j is the propagation operator for pipe j and Z_{cj} is the series impedance for pipe j, and are given by

$$\Gamma(s) = \sqrt{R_{oj} C_{oj} \left[s + R_j(s)\right] \left[s + C_j(s)\right]}, \quad Z_{cj}(s) = \sqrt{\left(\frac{R_{oj}}{C_{oj}}\right) \frac{s + R_j(s)}{s + C_j(s)}}$$

where R_j and C_j are the Laplace transforms of the linearised approximations of $\mathcal{R}$ and $\mathcal{C}$ respectively (typically the only term requiring linearisation is the steady-state quadratic term in $\mathcal{R}$, as for turbulent flow, $\mathcal{R}[q] = \overline{\mathcal{R}}[q] + O\left\{(q - q_o)^2\right\}$ where $q_o \neq 0$ is a reference flow rate (Wylie and Streeter 1993)). The propagation operator Γ_j characterises the amplitude and phase change of a propagating travelling wave, and Z_{cj} characterises the amplitude and phase coupling between the pressure and flow within a pipeline.

Given the node partitioning outlined in the previous section, $\boldsymbol{Y}$ can be expressed in the block matrix form

$$\boldsymbol{Y}(s) = \begin{pmatrix} \boldsymbol{Y}_{\mathcal{A}_1\mathcal{A}_1}(s) & \cdots & \boldsymbol{Y}_{\mathcal{A}_1\mathcal{C}_2}(s) \\ \vdots & \ddots & \vdots \\ \boldsymbol{Y}_{\mathcal{C}_2\mathcal{A}_1}(s) & \cdots & \boldsymbol{Y}_{\mathcal{C}_2\mathcal{C}_2}(s) \end{pmatrix} \tag{16}$$

where the block matrices $\boldsymbol{Y}_{AB}$ in $\boldsymbol{Y}$ are lexicographically ordered based on the pair (A, B) where $A, B \in \{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \mathcal{A}_4, \mathcal{B}_1, \mathcal{B}_2, \mathcal{C}_1, \mathcal{C}_2\}$. The matrices $\boldsymbol{Y}_{AB}$ are $n_A \times n_B$ matrices that can be interpreted to be the admittance mapping from $\boldsymbol{\Psi}_B$, the nodal pressures from set B, to $\boldsymbol{\Theta}_A$, the nodal flows for the nodes in set A.

Without loss of generality it can be assumed that the transforms of the known variables $\boldsymbol{\Psi}_k$ and $\boldsymbol{\Theta}_k$ are zero. This is a reasonable assumption as either the pressure is held constant (in the case of a reservoir) or the flow injection is zero (in the case of a junction) which means that there are no dynamic fluctuations in these variables about the steady-state point. Despite the fact that the method is formulated only for simple nodes (junctions and reservoirs), more complex boundary conditions can be incorporated into the framework by either treating the boundary conditions as unknown, or incorporating nodal variable measurements at these nodes. Retaining the important terms and collecting the measured and unmeasured variables yields the following expression of the network dynamics where the influences of the measured and unmeasured variables are made explicit

$$\boldsymbol{G}_m(s)\begin{pmatrix}\boldsymbol{\Psi}_m(s)\\ \boldsymbol{\Theta}_m(s)\end{pmatrix}+\boldsymbol{G}_u(s)\begin{pmatrix}\boldsymbol{\Psi}_u(s)\\ \boldsymbol{\Theta}_u(s)\end{pmatrix}=\boldsymbol{0} \tag{17}$$

and where the operators acting on the measured and unmeasured states are given by

$$\boldsymbol{G}_m(s)=\begin{pmatrix}\boldsymbol{Y}_{m1}(s) & -\boldsymbol{I}\\ \boldsymbol{Y}_{m2}(s) & \boldsymbol{0}\\ \boldsymbol{Y}_{m3}(s) & \boldsymbol{0}\end{pmatrix},\quad \boldsymbol{G}_u(s)=\begin{pmatrix}\boldsymbol{Y}_{u1}(s) & \boldsymbol{0}\\ \boldsymbol{Y}_{u2}(s) & \boldsymbol{0}\\ \boldsymbol{Y}_{u3}(s) & -\boldsymbol{I}\end{pmatrix}. \tag{18}$$

where the matrix transfer functions $\boldsymbol{Y}_{mi}$ $\boldsymbol{Y}_{ui}$, $i=1,2,3$, comprise the blocks in (16) as

$$\begin{aligned}
\boldsymbol{Y}_{m1}&=\begin{pmatrix}\boldsymbol{Y}_{\mathcal{A}_1\mathcal{A}_1} & \boldsymbol{Y}_{\mathcal{A}_1\mathcal{A}_2} & \boldsymbol{Y}_{\mathcal{A}_1\mathcal{B}_1}\\ \boldsymbol{Y}_{\mathcal{A}_3\mathcal{A}_1} & \boldsymbol{Y}_{\mathcal{A}_3\mathcal{A}_2} & \boldsymbol{Y}_{\mathcal{A}_3\mathcal{B}_1}\\ \boldsymbol{Y}_{\mathcal{C}_1\mathcal{A}_1} & \boldsymbol{Y}_{\mathcal{C}_1\mathcal{A}_2} & \boldsymbol{Y}_{\mathcal{C}_1\mathcal{B}_1}\end{pmatrix}, \quad \boldsymbol{Y}_{u1}=\begin{pmatrix}\boldsymbol{Y}_{\mathcal{A}_1\mathcal{A}_3} & \boldsymbol{Y}_{\mathcal{A}_1\mathcal{A}_4} & \boldsymbol{Y}_{\mathcal{A}_1\mathcal{B}_2}\\ \boldsymbol{Y}_{\mathcal{A}_3\mathcal{A}_3} & \boldsymbol{Y}_{\mathcal{A}_3\mathcal{A}_4} & \boldsymbol{Y}_{\mathcal{A}_3\mathcal{B}_2}\\ \boldsymbol{Y}_{\mathcal{C}_1\mathcal{A}_3} & \boldsymbol{Y}_{\mathcal{C}_1\mathcal{A}_4} & \boldsymbol{Y}_{\mathcal{C}_1\mathcal{B}_2}\end{pmatrix},\\
\boldsymbol{Y}_{m2}&=\begin{pmatrix}\boldsymbol{Y}_{\mathcal{B}_1\mathcal{A}_1} & \boldsymbol{Y}_{\mathcal{B}_1\mathcal{A}_2} & \boldsymbol{Y}_{\mathcal{B}_1\mathcal{B}_1}\\ \boldsymbol{Y}_{\mathcal{B}_2\mathcal{A}_1} & \boldsymbol{Y}_{\mathcal{B}_2\mathcal{A}_2} & \boldsymbol{Y}_{\mathcal{B}_2\mathcal{B}_1}\end{pmatrix}, \quad \boldsymbol{Y}_{u2}=\begin{pmatrix}\boldsymbol{Y}_{\mathcal{B}_1\mathcal{A}_3} & \boldsymbol{Y}_{\mathcal{B}_1\mathcal{A}_4} & \boldsymbol{Y}_{\mathcal{B}_1\mathcal{B}_2}\\ \boldsymbol{Y}_{\mathcal{B}_2\mathcal{A}_3} & \boldsymbol{Y}_{\mathcal{B}_2\mathcal{A}_4} & \boldsymbol{Y}_{\mathcal{B}_2\mathcal{B}_2}\end{pmatrix},\\
\boldsymbol{Y}_{m3}&=\begin{pmatrix}\boldsymbol{Y}_{\mathcal{A}_2\mathcal{A}_1} & \boldsymbol{Y}_{\mathcal{A}_2\mathcal{A}_2} & \boldsymbol{Y}_{\mathcal{A}_2\mathcal{B}_1}\\ \boldsymbol{Y}_{\mathcal{A}_4\mathcal{A}_1} & \boldsymbol{Y}_{\mathcal{A}_4\mathcal{A}_2} & \boldsymbol{Y}_{\mathcal{A}_4\mathcal{B}_1}\\ \boldsymbol{Y}_{\mathcal{C}_2\mathcal{A}_1} & \boldsymbol{Y}_{\mathcal{C}_2\mathcal{A}_2} & \boldsymbol{Y}_{\mathcal{C}_2\mathcal{B}_1}\end{pmatrix}, \quad \boldsymbol{Y}_{u3}=\begin{pmatrix}\boldsymbol{Y}_{\mathcal{A}_2\mathcal{A}_3} & \boldsymbol{Y}_{\mathcal{A}_2\mathcal{A}_4} & \boldsymbol{Y}_{\mathcal{A}_2\mathcal{B}_2}\\ \boldsymbol{Y}_{\mathcal{A}_4\mathcal{A}_3} & \boldsymbol{Y}_{\mathcal{A}_4\mathcal{A}_4} & \boldsymbol{Y}_{\mathcal{A}_4\mathcal{B}_2}\\ \boldsymbol{Y}_{\mathcal{C}_2\mathcal{A}_3} & \boldsymbol{Y}_{\mathcal{C}_2\mathcal{A}_4} & \boldsymbol{Y}_{\mathcal{C}_2\mathcal{B}_2}\end{pmatrix}.
\end{aligned} \tag{19}$$

In summary, (17) provides us with the basic model for considering the network dynamics as being dependent on both measured and unmeasured nodal states.

For systems where the links are passive (*i.e.*, they dissipate energy (Wohlers 1969)), these matrices have some properties that are important for the ensuing analysis. These are summarized in the following.

Theorem 1. *For a network* $(\mathcal{G}(\mathcal{N},\Lambda),\mathcal{P})$ *with a given nodal partitioning, if all links* $\lambda \in \Lambda$ *are passive, then the following relationships hold*

$$\boldsymbol{G}_u(s) \quad \textit{is full column rank,} \tag{20}$$

$$\begin{pmatrix} \boldsymbol{G}_m(s) & \boldsymbol{G}_u(s) \end{pmatrix} \quad \textit{is full row rank,} \tag{21}$$

for all $s \in \mathbb{C}_+$.

For brevity, the proof of this theorem is given in Appendix I, however, it is worth interpreting the meaning of these properties: (20) means that each unmeasured state influences the system dynamics in a way that is different from every other unmeasured state; and (21) can be interpreted to mean that each row in (17) describes a unique and linearly independent dynamic relationship between the network state variables.

Example. *For the example network from Figure 1(b) with the nodal partitioning as outlined as Scenario 1 in Table 2, the matrices from* (18) *are given as*

$$\boldsymbol{G}_m = \left[\begin{array}{cccc:c} Y_{44} & Y_{42} & Y_{43} & Y_{46} & -1 \\ \hdashline Y_{24} & Y_{22} & Y_{23} & Y_{26} & 0 \\ Y_{34} & Y_{32} & Y_{33} & Y_{36} & 0 \\ Y_{64} & Y_{62} & Y_{63} & Y_{66} & 0 \\ Y_{54} & Y_{52} & Y_{53} & Y_{56} & 0 \\ \hdashline Y_{74} & Y_{72} & Y_{73} & Y_{76} & 0 \\ Y_{14} & Y_{12} & Y_{13} & Y_{16} & 0 \end{array}\right], \boldsymbol{G}_u = \left[\begin{array}{cc:cc} Y_{47} & Y_{45} & 0 & 0 \\ \hdashline Y_{27} & Y_{25} & 0 & 0 \\ Y_{37} & Y_{35} & 0 & 0 \\ Y_{67} & Y_{65} & 0 & 0 \\ Y_{57} & Y_{55} & 0 & 0 \\ \hdashline Y_{77} & Y_{75} & -1 & 0 \\ Y_{17} & Y_{15} & 0 & -1 \end{array}\right] \tag{22}$$

where $Y_{ik} = Y_{ik}(s)$ *symbolises the* (i,k)*-th term in the network admittance matrix* $\boldsymbol{Y}(s)$ *from*

(16), and the partition lines indicate the submatrices in the matrix expressions from (18). From Theorem 1, the columns of $\boldsymbol{G}_u$ are linearly independent (on $s \in \mathbb{C}_+$) and represent the unique influence that each unmeasured nodal state $\boldsymbol{\psi}_u = [\psi_7\, \psi_5]^T$ and $\boldsymbol{\theta}_u = [\theta_7\, \theta_1]^T$ has on the system dynamics, and the rows of the matrix $(\boldsymbol{G}_m\, \boldsymbol{G}_u)$ are linearly independent (on $s \in \mathbb{C}_+$) indicating that eash row describes a unique dynamic relationship between the measured nodal states, $\boldsymbol{\psi}_m = [\psi_4\, \psi_2\, \psi_3\, \psi_6]^T$ and $\boldsymbol{\theta}_m = \theta_4$, and unmeasured nodal states.

THE PROPOSED EXPECTATION-MAXIMISATION ALGORITHM

To construct an error function on which to base a parameter estimate requires a model to describe the relationship between the measurements. The expression (17) provides us with such a description, however, it cannot be used in its present form due to the presence of the unmeasured terms $\boldsymbol{\Psi}_u$ and $\boldsymbol{\theta}_u$. In order to undertake an estimation procedure, this dependency must be accounted for. In the work by Zecchin et al. (2013), a decoupling filter $\boldsymbol{L}$ was constructed to nullify the influence of the unmeasured states on the system dynamics (*i.e.* $\boldsymbol{L}\boldsymbol{G}_u = \boldsymbol{0}$ for all $s \in \mathbb{C}_+$). However, this was only possible for the case where $\mathcal{A}_4 = \emptyset$, and so cannot be used for the more general case here. Instead, a different avenue is pursued through the application of the EM algorithm (Dempster et al. 1977), which provides a way of undertaing parameter estimation in systems involving *hidden* or unknown/unmeasured states.

Stated in its general form, given a system with measured states $\mathcal{U}_m$ and unmeasured states $\mathcal{U}_u$ drawn from the joint distribution $f\left(\mathcal{U}_m, \mathcal{U}_u | \boldsymbol{\vartheta}\right)$ parameterized by $\boldsymbol{\vartheta}$, for a given initial estimate $\boldsymbol{\vartheta}_0$ the following sequence of iterates

$$\boldsymbol{\vartheta}_k = \arg\max_{\boldsymbol{\vartheta}} \mathrm{E}\left[\ln\left(f\left(\mathcal{U}_m, \mathcal{U}_u | \boldsymbol{\vartheta}\right)\right) | \mathcal{U}_m, \boldsymbol{\vartheta}_{k-1}\right], \quad k = 1, \dots \tag{23}$$

converges to a local maximizer of the marginal likelihood function $f\left(\mathcal{U}_m | \boldsymbol{\vartheta}\right)$ of the measured states $\mathcal{U}_m$ (provided the marginal distribution is bounded). The process (23) has many different interpretations (Michiko and Kazunori 2004), but the most simple explanation is as follows: the k-th iterate $\boldsymbol{\vartheta}_k$ is given as the value that maximizes the expected log-likelihood of the joint distribution of the measured and unmeasured states $f\left(\mathcal{U}_m, \mathcal{U}_u | \boldsymbol{\vartheta}\right)$ over the conditional probability space $\mathcal{U}_u | \mathcal{U}_m, \boldsymbol{\vartheta}_{k-1}$ of the unmeasured states given the measured states and the $(k-1)$-th parameter estimate. The useful aspect of this approach is that the expectation integrates over the unmeasured variables explicitly removing them from the maximization function.

Expectation-maximisation for the $(\mathcal{G}(\mathcal{N}, \Lambda), \mathcal{P})$ network

Before the EM approach can be developed, the joint distribution f between the measured and unmeasured states must first be defined. For a network with a given nodal partitioning, in the case of a system in steady-oscillatory flow, the measurement data set comprises the time domain sequence (7) and (8) where $\boldsymbol{e}_\psi$ and $\boldsymbol{e}_\theta$ represent the measurement noise, and are stationary processes with power spectra $\boldsymbol{S}_\psi(\mathrm{i}\omega)$ and $\boldsymbol{S}_\theta(\mathrm{i}\omega)$. Given this form of the time-domain measurements for $N = 2M$, the frequency-domain data, as obtained through a discrete Fourier transform (DFT) (Brillinger 1974), follows the distribution

$$\begin{pmatrix} \widetilde{\boldsymbol{\Psi}}_m(\mathrm{i}\omega_i) \\ \widetilde{\boldsymbol{\Theta}}_m(\mathrm{i}\omega_i) \end{pmatrix} \sim \mathcal{N}_c\left(\begin{pmatrix} \boldsymbol{\Psi}_m(\mathrm{i}\omega_i) \\ \boldsymbol{\Theta}_m(\mathrm{i}\omega_i) \end{pmatrix}, \frac{1}{2M}\begin{pmatrix} \boldsymbol{S}_\psi(\mathrm{i}\omega_i) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{S}_\theta(\mathrm{i}\omega_i) \end{pmatrix}\right), \quad i = 1, \ldots, M \quad (24)$$

where the ω_i are the Fourier frequencies, and $\mathcal{N}_c$ is the complex normal distribution (Schoukens and Pintelon 1991). It is important to note that complex normal relationship (24) assumes only that the time-domain noise is a stationary process, where the system dynamics are embedded in the frequency dependent mean values of the data (the actual state values $\boldsymbol{\Psi}_m$ and $\boldsymbol{\Theta}_m$) as these correspond to the system's noise-free frequency response (i.e. no restriction is imposed on $\boldsymbol{\Psi}_m$ and $\boldsymbol{\Theta}_m$). The necessity of the EM algorithm comes into play because the mean values of the measured data $\boldsymbol{\Psi}_m$ and $\boldsymbol{\Theta}_m$ are unknown, but are dependent on the unmeasured nodal states $\boldsymbol{\Psi}_u$ and $\boldsymbol{\Theta}_u$ through (17). Therefore, the total data that comprises the measured and unmeasured nodal states must be considered. To determine the joint distribution of the total data, it is required to assign a distribution to the unmeasured data.

Therefore, assuming that the data for the unmeasured states follows

$$\begin{pmatrix} \widetilde{\boldsymbol{\Psi}}_u(\mathrm{i}\omega_i) \\ \widetilde{\boldsymbol{\Theta}}_u(\mathrm{i}\omega_i) \end{pmatrix} \sim \mathcal{N}_c \left(\begin{pmatrix} \boldsymbol{\Psi}_u(\mathrm{i}\omega_i) \\ \boldsymbol{\Theta}_u(\mathrm{i}\omega_i) \end{pmatrix}, \boldsymbol{A}_i \right), \quad i = 1, \ldots, M, \tag{25}$$

where $\boldsymbol{A}_i$ is a symmetric positive definite matrix, the joint distribution of the total data is given by (24) and (25) where the means are unknown, but related by (17).

Now that we have a joint distribution between our measured and unmeasured states, the EM algorithm can be applied. For simplicity, the majority of the analysis is deferred to Appendix II, but the main results concerning the final EM algorithm are summarized in the following theorem.

Theorem 2. *Consider a $(\mathcal{G}(\mathcal{N}, \Lambda), \mathcal{P})$ network with passive links, a given nodal partitioning, and the measured and unmeasured data sets $\widetilde{\mathcal{U}} = \{\widetilde{\boldsymbol{u}}_1, \ldots, \widetilde{\boldsymbol{u}}_M\}$ and $\widetilde{\mathcal{V}} = \{\widetilde{\boldsymbol{v}}_1, \ldots, \widetilde{\boldsymbol{v}}_M\}$, respectively, where*

$$\widetilde{\boldsymbol{u}}_i = \begin{pmatrix} \widetilde{\boldsymbol{\Psi}}_m(\mathrm{i}\omega_i) \\ \widetilde{\boldsymbol{\Theta}}_m(\mathrm{i}\omega_i) \end{pmatrix}, \quad \widetilde{\boldsymbol{v}}_i = \begin{pmatrix} \widetilde{\boldsymbol{\Psi}}_u(\mathrm{i}\omega_i) \\ \widetilde{\boldsymbol{\Theta}}_u(\mathrm{i}\omega_i) \end{pmatrix}$$

are distributed as in (24) *and* (25) *with mean values, $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$, related by the system equation*

$$\begin{pmatrix} \boldsymbol{G}_{mi}(\boldsymbol{\vartheta}) & \boldsymbol{G}_{ui}(\boldsymbol{\vartheta}) \end{pmatrix} \begin{pmatrix} \boldsymbol{u}_i \\ \boldsymbol{v}_i \end{pmatrix} = \boldsymbol{0}, \quad i = 1, \ldots, M \tag{26}$$

where $\boldsymbol{G}_{mi} = \boldsymbol{G}_m(\mathrm{i}\omega_i)$ and $\boldsymbol{G}_{ui} = \boldsymbol{G}_u(\mathrm{i}\omega_i)$ are parameterized by the network parameter set $\boldsymbol{\vartheta} \in \boldsymbol{\Upsilon}$ [$\boldsymbol{G}_m$ and $\boldsymbol{G}_u$ are the system matrices from (18)*] . The EM algorithm for estimating $\boldsymbol{\vartheta}$ based on the measured data $\widetilde{\mathcal{U}}$ is given by the sequence*

$$\boldsymbol{\vartheta}_{k+1} = \arg\max_{\boldsymbol{\vartheta} \in \boldsymbol{\Upsilon}} - \sum_{i=1}^{M} Q_i \left(\widetilde{\boldsymbol{u}}_i, \boldsymbol{\vartheta}_k, \boldsymbol{\vartheta} \right) \tag{27}$$

where $Q_i \left(\widetilde{\boldsymbol{u}}_i, \boldsymbol{\vartheta}_k, \boldsymbol{\vartheta} \right)$ is the negative of the expectation of the log likelihood of the joint dis-

tribution of $\widetilde{\boldsymbol{u}}_i$ and $\widetilde{\boldsymbol{v}}_i$ conditional on $\widetilde{\boldsymbol{u}}_i$ and $\boldsymbol{\vartheta}_k$ (constant terms neglected), and is given by

$$
\begin{aligned}
Q_i\left(\widetilde{\boldsymbol{u}}, \boldsymbol{\vartheta}_k, \boldsymbol{\vartheta}\right) = \; & \widetilde{\boldsymbol{u}}^H \boldsymbol{C}_{mmi}\left(\boldsymbol{\vartheta}\right) \widetilde{\boldsymbol{u}} \\
& -2\mathbb{R}\mathrm{e}\left\{\widetilde{\boldsymbol{u}}^H \boldsymbol{C}_{mui}\left(\boldsymbol{\vartheta}\right) \boldsymbol{C}_{uui}^{-1}\left(\boldsymbol{\vartheta}_k\right) \boldsymbol{C}_{umi}\left(\boldsymbol{\vartheta}_k\right) \widetilde{\boldsymbol{u}}\right\} \\
& +\mathrm{tr}\left\{\boldsymbol{C}_{uui}\left(\boldsymbol{\vartheta}\right) \boldsymbol{\Sigma}_{ui}\right\} \\
& +\widetilde{\boldsymbol{u}}^H \boldsymbol{C}_{mui}\left(\boldsymbol{\vartheta}_k\right) \boldsymbol{C}_{uui}^{-1}\left(\boldsymbol{\vartheta}_k\right) \boldsymbol{C}_{uui}\left(\boldsymbol{\vartheta}\right) \boldsymbol{C}_{uui}^{-1}\left(\boldsymbol{\vartheta}_k\right) \boldsymbol{C}_{umi}\left(\boldsymbol{\vartheta}_k\right) \widetilde{\boldsymbol{u}}
\end{aligned} \quad . \tag{28}
$$

where

$$
\left.\begin{aligned}
\boldsymbol{C}_{mmi}(\boldsymbol{\vartheta}) &= \boldsymbol{G}_{mi}^H(\boldsymbol{\vartheta})\boldsymbol{\Lambda}_i^{-1}(\boldsymbol{\vartheta})\boldsymbol{G}_{mi}(\boldsymbol{\vartheta}) \\
\boldsymbol{C}_{mui}(\boldsymbol{\vartheta}) &= \boldsymbol{G}_{mi}^H(\boldsymbol{\vartheta})\boldsymbol{\Lambda}_i^{-1}(\boldsymbol{\vartheta})\boldsymbol{G}_{ui}(\boldsymbol{\vartheta}) \\
\boldsymbol{C}_{umi}(\boldsymbol{\vartheta}) &= \boldsymbol{G}_{ui}^H(\boldsymbol{\vartheta})\boldsymbol{\Lambda}_i^{-1}(\boldsymbol{\vartheta})\boldsymbol{G}_{mi}(\boldsymbol{\vartheta}) \\
\boldsymbol{C}_{uui}(\boldsymbol{\vartheta}) &= \boldsymbol{G}_{ui}^H(\boldsymbol{\vartheta})\boldsymbol{\Lambda}_i^{-1}(\boldsymbol{\vartheta})\boldsymbol{G}_{ui}(\boldsymbol{\vartheta})
\end{aligned}\right\} \tag{29}
$$

and

$$
\boldsymbol{\Lambda}_i(\boldsymbol{\vartheta}) = \boldsymbol{G}_{mi}(\boldsymbol{\vartheta})\boldsymbol{\Sigma}_{mi}\boldsymbol{G}_{mi}^H(\boldsymbol{\vartheta}) + \boldsymbol{G}_{ui}(\boldsymbol{\vartheta})\boldsymbol{\Sigma}_{ui}\boldsymbol{G}_{ui}^H(\boldsymbol{\vartheta}), \tag{30}
$$

where $\boldsymbol{\Sigma}_{mi}$ and $\boldsymbol{\Sigma}_{ui}$ are the covariance matrices for the measured and unmeasured data as in (24) *and* (25)*.*

A constructive proof of Theorem 2 is given in Appendix II. Within this theorem, it is seen that the EM algorithm resolves down to solving the sequence of maximization problems (27) to achieve increasingly accurate parameter estimates as k increases.

Computational Algorithm

An algorithm for computing the EM parameter estimate from Theorem 2 is outlined in Algorithm 1. The required input data to compute the EM parameter estimate is the network topology $\mathcal{G}\left(\mathcal{N}, \Lambda\right)$, frequency-domain data $\widetilde{\mathcal{U}}$ (corresponding frequencies $\boldsymbol{\omega} = \{\omega_1, \ldots, \omega_M\}$), covariances for measured (and unmeasured) data $\boldsymbol{\Sigma}_m$ and $\boldsymbol{\Sigma}_u$, a specified parameter range $\boldsymbol{\Upsilon}$, and an initial parameter estimate $\boldsymbol{\vartheta}_0$. As seen in Steps 1 and 2 of Algorithm 1, the first step involves determining the nodal sets $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ from Table 1, and reordering the nodal states as in (10). As outlined in Steps 3 to 10, with the initial parameter estimate $\boldsymbol{\vartheta}_0$, successive parameter estimates are determined as the maximiser of the expected conditional loglikelihood function from (27) (represented in Algorithm 1 as the function `ExpCondLogLikelihood`). The computational algorithm for `ExpCondLogLikelihood` is outlined in Algorithm 2 and discussed below. Once the termination criteria is met (typically a limit on the maximum k, or a lower threshold for the update norm $||\boldsymbol{\vartheta}_{k+1} - \boldsymbol{\vartheta}_k||$), the latest parameter estimate is returned as the EM estimate.

The crux of Algorithm 1 is the maximisation of the expected conditional loglikelihood function in Step 5 (equation (27)). Given the complexity of the optimisation problem, iterative techniques are necessary which require repeated calls to `ExpCondLogLikelihood` to determine the maximiser. The computation of `ExpCondLogLikelihood` is outlined in Algorithm 2 where the required inputs for Algorithm 2 follow those required for Algorithm 1. The algorithm loops through all frequencies $\omega_i = 1, \ldots, M$ (Steps 2 to 12), summating the expected conditional loglikelihood Q_i terms. At each frequency, the system matrices required to determine Q_i are computed for both the old parameter estimate $\boldsymbol{\vartheta}_k$ and the new parameter estimate $\boldsymbol{\vartheta}$ (Steps 3 to 9). For each parameter value, first the network admittance matrix is determined (Step 4), followed by the sequential construction and calculation of the system matrices (Steps 5 to 8). Once all the system matrices for both the old and new parameter estimates are determined, the Q_i term for the i-th frequency can be computed from (28) (Step 10). The Q_i terms are summated for each frequency, and finally the expected conditional loglikelihood function $-Q$ is returned (Step 13).

NUMERICAL EXAMPLES

The following numerical examples demonstrate the ability of the proposed EM algorithm from Theorem 2 to accurately estimate a hydraulic network's parameters in the presence of unknown boundary conditions. The 11-pipe network depicted in Figure 1 (b) is the focus of the study. The additional subnetwork comprising pipes [12] and [13] and nodes 8 and 9, in

Figure 1 (a), is treated as an unknown subnetwork, resulting in an unknown and unmeasured boundary condition at node 7.

Two different case studies, each with a different resistance function $\mathcal{R}$ are considered. For these case studies, the measured information consists of pressure measurements at nodes $\{2, 3, 4, 6\}$, and a flow measurement at node 4. In order to test the utility of the proposed EM algorithm for dealing with systems with unknown boundary conditions, two different scenarios of prior information were considered: Scenario 1, the existence of the additional connection at node 7 is known, hence an unknown flow boundary condition at node 7 is assumed (the correct assumption); Scenario 2, the existence of the additional connection at node 7 is not known, hence the flow boundary condition at node 7 is assumed to be zero (the incorrect assumption). These two scenarios allow for a direct testing of the effectiveness that the EM algorithm is able to deal with the unknown boundary condition. The nodal set partitions (from Table 1) corresponding to these scenarios is given in Table 2.

Preliminaries

The raw time-domain data for the numerical experiments was generated from a method of characteristics (MOC) simulation with added Gaussian noise. The frequency-domain data was obtained from the DFT of the time-domain data. For the MOC simulation, the system was excited into a steady-oscillatory transient state by a multi-sine flow perturbation at node 4 consisting of 983 equi-spaced frequencies from 0 to 15 Hz with amplitudes ranging from 0.01 to 0.1 L/s. The time-domain measurement errors were taken as independent zero mean Gaussian variates with standard deviations of 1 kPa for the pressure measurements and 0.32 L/s for the flow measurement. All results presented are based on 10 independent trial data sets.

For the purposes of the EM algorithm, the pipeline parameter values were assumed to be known down to an interval, where: the wavespeed c was known to be within [900,1200] m/s; the friction factor f within [0.015,0.04]; the diameters D were known to within $\pm$10 mm of their actual value; the pipe lengths l were known to within $\pm$20 m of their actual value; the relative roughnesses ϵ/D were known to be on the interval [0.0001, 0.01]; and the steady-state velocities v_o were known to be on the interval [0.1, 10] m/s. Narrowing the parameter values down to such intervals is reflective of the *a priori* knowledge available within real systems. Within the experiments, two different pipeline resistance functions were considered, namely the turbulent-steady-friction (TSF) model (Wylie and Streeter 1993), and the turbulent-unsteady-friction (TUF) model (Vardy and Brown 2007). For all experiments, the pipeline's were modelled elastically (*i.e.* $\mathcal{C} = 0$).

Algorithms 1 and 2 were used as the framework to determine the EM estimate, where the maximiser $\widehat{\boldsymbol{\vartheta}}$ was computed using the evolutionary algorithm process of particle swarm optimisation (PSO) (i.e. PSO was used to solve the optimisation problem in Step 5 of Algorithm 1). More details on the adopted optimization process are given in Zecchin (2010).

Case Study 1: Turbulent steady friction pipeline model

Pipeline model

Within the first case study, the network pipelines were modelled within the MOC using the TSF resistance function given by

$$\mathcal{R}_{\mathrm{TSF}}[q](x,t) = \frac{f}{2DA}|q(x,t)|q(x,t)$$

where f is the Darcy-Weisbach friction factor (Wylie and Streeter 1993), and A and D are the cross-sectional area and pipe diameter. As the proposed frequency-domain method is a linear approximation, the EM algorithm assumed a resistance function of the form $\overline{\mathcal{R}}_{\mathrm{TSF}}[q](x,t) = (fv_o/D)q(x,t)$ (where v_o is the steady-state velocity), leading to the following expressions for the frequency-domain pipeline functions Γ and Z_c

$$\Gamma(s) = \Gamma_o\sqrt{s\,(s + r_o)}, \quad Z_c(s) = Z_{co}\sqrt{\frac{s + r_o}{s}}$$

where

$$\Gamma_o = \sqrt{R_o C_o} = \frac{l}{c}, \quad Z_{co} = \sqrt{\frac{R_o}{C_o}} = \rho\frac{c}{A}, \quad r_o = \frac{f v_o}{D} \tag{31}$$

where c is the wavespeed, l is the length, and ρ is the fluid density. The functions Γ and Z_c are dependent on five parameters c, D, l, f and $v_o \neq 0$ (assuming that the density is known), however they only appear as the three terms (31). Therefore, the functions Γ and Z_c are described by the values of these three terms meaning a unique parameter set for the TSF pipeline model is $\boldsymbol{\vartheta} = \{\Gamma_o, Z_{co}, r_o\}$. Consequently, for the 11-pipe network, the parameter space to be identified is $\boldsymbol{\vartheta} = \{r_{o1}, \Gamma_{o1}, Z_{co1}\} \cup \cdots \cup \{r_{o11}, \Gamma_{o11}, Z_{co11}\}$ which is a total of 33-dimensions.

Results

The parameter estimation results for each of the prior information scenarios from Table 2 are summarised in the box plots in Figure 2 and the statistics in Table 3, where the relative error is defined as the difference between the estimated and actual parameter values, as a percentage of the actual parameter value. Comparing the performance of Scenario 1 to Scenario 2 from Figure 2 and Table 3, it is clear that the correct hypothesis concerning the node 7 flow (i.e. Scenario 1 that assumed an unknown nodal flow $\theta_7 \neq 0$) on average yielded more accurate parameter estimates than the incorrect hypothesis (i.e. Scenario 2 that assumed a known nodal flow of $\theta_7 = 0$). In Table 3 it is seen that the median error estimates for Scenario 1 are all lower than Scenario 2, this is particularly so for the estimates for r_o and Z_{co}. Most notably is that the error for the parameter estimate of r_o for pipe [10] was less than 5% for Scenario 1 but in the order of $O\{10^3\}$% for Scenario 2.

A more thorough consideration of Figure 2 shows that some stronger patterns exist within the data. The parameter estimates for Scenario 1 are significantly better than those for Scenario 2 for all pipes that are incident to node 7 (*i.e.* pipes [8], [9], and [10]). This pattern indicates two observations. Firstly, incorrect nodal categorisations have a more significant impact on the parameter estimates for links that are incident to nodes that have been incorrectly characterised. Secondly, the proposed EM algorithm has successfully provided accurate parameter estimates for a system containing a node for which no information exists (i.e. both the nodal pressure and flow are unknown and unmeasured). This has not been achieved before within the literature, to the authors knowledge.

Considering the Scenario 1 estimates for the different parameter types, it is observed that the propagation coefficient Γ_o is estimated with a high accuracy, far higher than the resistance coefficient r_o and the impedance coefficient Z_{co}. The hypothesised reason for this lies in the influence that the parameters have in the pattern of the system's frequency response. The parameter Γ_o is related to the period of a pipeline and hence the location of the harmonics in the frequency-domain, whereas r_o and Z_{co} are related to the energy dissipation within a pipeline and are hence related to the harmonic amplitudes in the frequency-domain. The error between the model predictions and the data is much more sensitive to mis-aligned harmonics than it is to well aligned harmonics with slightly different amplitudes. Therefore, by implication, it is expected that the error between the model predictions and the data would be much more sensitive to errors in the estimation of Γ_o compared to that of r_o and Z_{co}, resulting in more accurate estimates for Γ_o in comparison to r_o and Z_{co}.

This reasoning also explains why the parameter estimates of Γ_o for Scenario 2 were reasonably accurate despite the incorrect assumption about the flow at node 7. The presence of the branch from node 7 did not alter the locations of the network's harmonics that were associated with the periods of the known 11 pipes. Hence the Γ_o parameters were still able to be estimated accurately. However, the presence of the branch did serve to dissipate energy within the system through the combined action of pipe friction and losses through the emitter. Therefore, as the branch changed the network's harmonic amplitudes, the estimates for r_o and Z_{co} were affected as they are related to these amplitudes.

Case Study 2: Turbulent unsteady friction pipeline model

Pipeline model

Within the second case study, the experiments of the first case study were repeated using a different pipeline resistance function model, namely the TUF model given by Vardy and Brown (2007)

$$\mathcal{R}_{\mathrm{TUF}}[q](x,t) = \mathcal{R}_{\mathrm{TSF}}[q](x,t) + \int_0^t r(t-\tau)\frac{\partial q}{\partial t}(x,\tau)\mathrm{d}\tau$$

where r is a weighting function that is parametrically dependent on the pipe diameter D, the kinematic viscosity ν, the Reynolds number $\mathbb{R}_{\mathrm{e}} = v_o D/\nu$, and the relative roughness ϵ/D (see Vardy and Brown (2007) for details). The difference between the TSF and TUF models are that the TUF model accounts for the additional dissipation within the fluid body resulting from accelerating and decelerating flows. For the TUF model, the propagation operator and characteristic impedance are given by

$$\Gamma(s) = \Gamma_o\sqrt{s\left(s + r_o + r(s)\right)}, \quad Z_c(s) = Z_{co}\sqrt{\frac{s + r_o + r(s)}{s}}$$

where Γ_o, Z_{co}, and r_o are as defined above, but with f_o as a function of ϵ/D and $\mathbb{R}_{\mathrm{e}}$, and $r(s)$ is the Laplace transform of $r(t)$. Given that the fluid density and viscosity are known, Γ and Z_c can be uniquely parameterised by the parameter set $\boldsymbol{\vartheta} = \{c, D, l, C_\epsilon, C_{\mathbb{R}_{\mathrm{e}}}\}$ where $C_\epsilon = \log_{10}(\epsilon/D)$ and $C_{\mathbb{R}_{\mathrm{e}}} = \log_{10}\mathbb{R}_{\mathrm{e}}$. Consequently, the 11-pipe network parameter estimation problem for case study 2 involves the estimation of the 55 dimensional parameter set $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}_1 \cup \cdots \cup \boldsymbol{\vartheta}_{11}$ where $\boldsymbol{\vartheta}_i = \{c_i, D_i, l_i, C_{\epsilon i}, C_{\mathbb{R}_{\mathrm{e}} i}\}$.

Results

The results of 10 independent trials for each scenario are summarised in the box plots in Figure 3 and the statistics in Table 4. As demonstrated in Table 4, for Scenario 1 (the correct assumption about the node 7 flow), the EM algorithm on average yielded more accurate parameter estimates for all parameters except the pipe diameters. Consistent with case study 1, is the pattern that the parameter estimates for Scenario 1 are significantly better than those for Scenario 2 for all pipes that are incident to node 7 (*i.e.* pipes [8], [9], and [10]). This reinforces the observation that (i) the correct categorisation of a node is particularly crucial for the accurate estimation of the parameters of all links incident to that node, and (ii) the proposed EM algorithm is effectively able to deal with nodes for which there is no information (*i.e.* the transient fluctuations in nodal pressure and nodal flow are unknown and unmeasured).

Drawing from both case studies 1 and 2, more detail can be given to these conclusions, in that it is mainly the parameters associated with energy dissipation that are affected by the incorrect categorization of node 7. As Scenario 2 does not allow for any flow to leave node 7, the energy that enters links [8] to [10] is considered as only being dissipated within the links on the known 11-pipe network. The implication of this is that the energy loss parameters (*e.g.* r_o for TSF pipes and $C_{\mathbb{R}_e}$ for TUF pipes), will be higher than the actual values. However, as Scenario 1 correctly categorises node 7 and allows for energy loss through this node (through the correct categorisation of this node, allowing for unmeasured flow into and out of this node), the energy loss parameter estimates for the links connected to this node are more accurate.

As with all the parameter estimation examples within this paper, the variables related to the system harmonic locations (*i.e.* wavespeed and pipe length) were estimated with greater accuracy than the other parameters. This is particularly true for C_ϵ, where the apparent lack of sensitivity of the methodology to this parameter is attributed to the fact that ϵ/D only appears in the expression for the pipeline functions Γ and Z_c through the functions A^* and B^* (Vardy and Brown 2007), thus potentially diminishing its influence.

CONCLUSIONS

This paper presents a novel method for the estimation of hydraulic network parameters based on transient fluid state measurements using the expectation-maximisation (EM) method. The proposed method was formulated to deal with a broader class of measurement scenarios than has previously been considered within the literature, specifically, it is designed to deal with scenarios for which there exist unknown boundary conditions.

Within the proposed approach, the measured nodal states were treated as being only part of the complete data (the complete data consisting of both the measured and unmeasured nodal states). Based on posing the problem in a constrained complex Gaussian framework, the statistical EM algorithm was used to derive a scheme to estimate the network parameters based on only using the information from the measured nodal states. This proposed method is significant in that it is the only method within the literature that is able to deal with the case where there are nodes within the system for which no information exists.

A series of numerical experiments were performed by coupling the EM method with a particle swarm optimisation (PSO) algorithm. The experiments were designed to test the ability of the methodology to deal with unknown nodal states. This was undertaken by dealing with a 13-pipe network for which full topology of the network was considered unknown. That is, for the purposes of parameter identification, the network was considered as an 11-pipe network with an unknown nodal pressure and flow at one of the network nodes. The results indicated that the use of the EM approach to correctly deal with the unknown nodal variables resulted in parameter estimates of a greater accuracy, particularly for the parameters of pipes incident to nodes for which no information exists.

The proposed method provides a small step closer to dealing with parameter identification in realistic networks by providing a statistically posed way of dealing with uncertain boundary conditions. However, within realworld networks, there still remains many problems to be solved such as uncertainties in the system dynamics, noise present within the system from external sources, and the incorporation of complex boundary conditions (such as actively controlled pumps and control valves).

ACKNOWLEDGEMENTS

This research has been financially supported by the Australian Research Council.

REFERENCES

Brillinger, D. R. (1974). *Time series: data analysis and theory*. International series in decision processes. Holt Rinehart and Winston, New York.

Dempster, A. P., Larid, N. M., and Rubin, D. B. (1977). "Maximum likelihood from incomplete data set via the EM algorithm." *Journal of the Royal Statistical Society.*, B 39(1), 1–38.

Diestel, R. (2000). *Graph Theory*. Springer-Verlag, New York, USA, electronic edition edition.

Isermann, R. (1984). "Process fault-detection based on modeling and estimation methods - a survey." *Automatica*, 20(4), 387–404.

Kim, S. H. (2008). "Address-oriented impedance matrix method for generic calibration of heterogeneous pipe network systems." *Journal of Hydraulic Engineering, ASCE*, 134(1), 66–75.

Lee, P. J., Vítkovský, J. P., Lambert, M. F., Simpson, A. R., and Liggett, J. A. (2005). "Frequency domain analysis for detecting pipeline leaks." *Journal of Hydraulic Engineering, ASCE*, 131(7), 596–604.

Liggett, J. A. and Chen, L. C. (1994). "Inverse transient analysis in pipe networks." *Journal of Hydraulic Engineering, ASCE*, 120(8), 934–955.

Liou, J. C. P. and Tian, J. (1995). "Leak detection - transient flow simulation approaches." *Journal of Energy Resources Technology*, 117(September), 243–248.

Mathai, A. M. and Provost, S. B. (1992). *Quadratic forms in random variables : theory and applications*, Vol. 126 of *Statistics, textbooks and monographs*. M. Dekker, New York.

W. Michiko and Y. Kazunori, eds. (2004). *The EM algorithm and related statistical models*, Vol. 170 of *Statistics, textbooks and monographs*. Marcel Dekker, New York.

Mohapatra, P. K., Chaudhry, M. H., Kassem, A., and Moloo, J. (2006). "Detection of partial blockages in a branched piping system by the frequency response method." *Journal of Fluids Engineering-Transactions of the ASME*, 128(5), 1106–1114.

Nash, G. A. and Karney, B. W. (1999). "Efficient inverse transient analysis in series pipe systems." *Journal of Hydraulic Engineering, ASCE*, 125(7, July), 761–764.

Rice, J. A. (1995). *Mathematical statistics and data analysis*. Wadsworth, Belmont, CA, 2nd edition.

Rieutord, E. and Blanchard, A. (1979). "Pulsating viscoelastic pipe flow - water-hammer." *Journal of Hydraulic Research, IAHR*, 17(3), 217–229. in French.

Schoukens, J. and Pintelon, R. (1991). *Identification of linear systems : a practical guideline to accurate modeling*. Pergamon Press, Oxford ; New York, 1st edition.

Stecki, J. S. and Davis, D. C. (1986). "Fluid transmission-lines - distributed parameter models . 1. A review of the state-of-the-art." *Proceedings of the Institution of Mechanical Engineers Part A-Journal of Power and Energy*, 200(4), 215–228.

Vardy, A. E. and Brown, J. M. B. (2007). "Approximation of turbulent wall shear stresses in highly transient pipe flows." *Journal of Hydraulic Engineering-ASCE*, 133(11), 1219–1228.

Verde, C., Visairo, N., and Gentil, S. (2007). "Two leaks isolation in a pipeline by transient response." *Advances in Water Resources*, 30(8), 1711–1721.

Wang, X. J., Lambert, M. F., Simpson, A. R., Liggett, J. A., and Vitkovsky, J. P. (2002). "Leak detection in pipelines using the damping of fluid transients." *Journal of Hydraulic Engineering, ASCE*, 128(7), 697–711.

Wohlers, M. R. (1969). *Lumped and Distributed passive networks; a generalised and advanced viewpoint*. Electrical science. Academic Press, New York.

Wylie, E. B. and Streeter, V. L. (1993). *Fluid Transients in Systems*. Prentice-Hall Inc., Englewood Cliffs, New Jersey, USA.

Zecchin, A. C. (2010). "Laplace-domain analysis of fluid line networks with applications to time-domain simulation and system parameter identification," PhD, The School of Civil, Environmental and Mining Engineering, The University of Adelaide.

Zecchin, A. C., Lambert, M. F., and Simpson, A. R. (2012). "Inverse laplace transform for transient-state fluid line network simulation." *Journal of Engineering Mechanics-ASCE*, 138(1), 101–115.

Zecchin, A. C., Lambert, M. F., Simpson, A. R., and White, L. B. (2010). "Frequency-domain modeling of transients in pipe networks with compound nodes using a Laplace-domain admittance matrix." *Journal of Hydraulic Engineering, ASCE*, 136(10), 739–755.

Zecchin, A. C., Simpson, A. R., Lambert, M. F., White, L. B., and Vitkovsky, J. P. (2009). "Transient modeling of arbitrary pipe networks by a Laplace-domain admittance matrix." *Journal of Engineering Mechanics, ASCE*, 135(6), 538–547.

Zecchin, A. C., White, L. B., Lambert, M. F., and Simpson, A. R. (2013). "Parameter identification of fluid line networks by frequency-domain maximum likelihood estimation." *Mechanical Systems and Signal Processing*, 37(1-2).

APPENDIX I. PROOF OF THEOREM 1

First considering (20). Given (18), $\boldsymbol{G}_u$ is full column rank when the first block column is full column rank. The rows and columns of $\boldsymbol{G}_u$ can be reordered to show that a principal minor of $\boldsymbol{Y}$ is embedded within this first block row, which is known to be positive definite on $s \in \mathbb{C}_+$ for a network comprised of passive links (see Zecchin (2010) for details). Therefore, as this first block row contains a positive definite submatrix, it is full column rank.

Now considering (21). Given the structure of the identities in both $\boldsymbol{G}_m$ and $\boldsymbol{G}_u$ in (18), it is clear that both the top and bottom block rows of $(\boldsymbol{G}_m \ \boldsymbol{G}_u)$ are linearly independent for all $s \in \mathbb{C}_+$. Therefore, to demonstrate the full row rank nature of this matrix, it is required to show that the center block row is a full row rank matrix. Reordering the columns and rows of $(\boldsymbol{G}_m \ \boldsymbol{G}_u)$, it can be shown that a principal minor of $\boldsymbol{Y}$ is embedded in the center block row. As stated above, this submatrix is positive definite on $s \in \mathbb{C}_+$ meaning that the center block row is full row rank (see Zecchin (2010) for details).

APPENDIX II. PROOF OF THEOREM 2

The determination of an EM process for the parameter $\boldsymbol{\vartheta}$ requires three distinct steps. Firstly, the determination of the joint probability density function (PDF) for the $\widetilde{\boldsymbol{u}}_i$ and $\widetilde{\boldsymbol{v}}_i$ given analytic forms of the maximum likelihood estimation (MLE) for the unknown means $\widehat{\boldsymbol{u}}_i$ and $\widehat{\boldsymbol{v}}_i$. Secondly, the expectation of the log-likelihood of the joint density over the conditional density of the unmeasured data $\widetilde{\boldsymbol{v}}_i$ given our measured data $\widetilde{\boldsymbol{u}}_i$ and $\boldsymbol{\vartheta}$. Thirdly, the expression of this log-likelihood purely as a function of $\widetilde{\boldsymbol{u}}_i$ by determining an estimate for $\boldsymbol{v}_i$ given only $\widetilde{\boldsymbol{u}}_i$.

Concerning the first step, as all the variates are independent for each $i = 1, \ldots, M$, the analytic expression of the MLEs for the unknown means $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ as a function of $\widetilde{\boldsymbol{v}}_i$, $\widetilde{\boldsymbol{u}}_i$, and $\boldsymbol{\vartheta}$ are given as (Zecchin 2010)

$$\begin{pmatrix} \widehat{\boldsymbol{u}}_i(\boldsymbol{\vartheta}) \\ \widehat{\boldsymbol{v}}_i(\boldsymbol{\vartheta}) \end{pmatrix} = \left(\boldsymbol{I} - \begin{pmatrix} \boldsymbol{\Sigma}_{mi} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_{ui} \end{pmatrix} \begin{pmatrix} \boldsymbol{G}_{mi}^H \\ \boldsymbol{G}_{ui}^H \end{pmatrix} \boldsymbol{\Lambda}_i^{-1} \begin{pmatrix} \boldsymbol{G}_{mi} & \boldsymbol{G}_{ui} \end{pmatrix} \right) \begin{pmatrix} \widetilde{\boldsymbol{u}}_i \\ \widetilde{\boldsymbol{v}}_i \end{pmatrix} \tag{32}$$

where $\boldsymbol{\Lambda}_i$ is as given in (30). The existence of $\boldsymbol{\Lambda}_i^{-1}$is ensured by the positive definiteness of $\boldsymbol{\Lambda}_i$ arising from the fact that diag $\{\boldsymbol{\Sigma}_{mi}, \boldsymbol{\Sigma}_{mi}\}$ is positive definite, and $(\boldsymbol{G}_{mi}\ \boldsymbol{G}_{ui})$ is full row rank (Theorem 1). Given these MLEs, the joint distribution for the measured and unmeasured i-th vector variates is

$$f_i\left(\widetilde{\boldsymbol{u}}_i, \widetilde{\boldsymbol{v}}_i | \widehat{\boldsymbol{u}}_i, \widehat{\boldsymbol{v}}_i, \boldsymbol{\vartheta}\right) = \frac{1}{\pi^n |\boldsymbol{\Sigma}_{mi}||\boldsymbol{\Sigma}_{ui}|} \exp\left\{ - \begin{pmatrix} \widetilde{\boldsymbol{u}}_i \\ \widetilde{\boldsymbol{v}}_i \end{pmatrix}^H \begin{pmatrix} \boldsymbol{C}_{mmi}(\boldsymbol{\vartheta}) & \boldsymbol{C}_{mui}(\boldsymbol{\vartheta}) \\ \boldsymbol{C}_{umi}(\boldsymbol{\vartheta}) & \boldsymbol{C}_{uui}(\boldsymbol{\vartheta}) \end{pmatrix} \begin{pmatrix} \widetilde{\boldsymbol{u}}_i \\ \widetilde{\boldsymbol{v}}_i \end{pmatrix} \right\} \tag{33}$$

for $i = 1, \ldots, M$, where the complex matrix functions $\boldsymbol{C}_{mmi}$, $\boldsymbol{C}_{mui}$, $\boldsymbol{C}_{mmi}$, and $\boldsymbol{C}_{uui}$ are given by (29). Finally, the negative of the log-likelihood of the joint distribution (33) can be expressed as

$$
\begin{aligned}
&-\ln\left(f_i\left(\widetilde{\boldsymbol{u}}_i, \widetilde{\boldsymbol{v}}_i | \widehat{\boldsymbol{u}}_i, \widehat{\boldsymbol{v}}_i, \boldsymbol{\vartheta}\right)\right) = \\
&\quad \pi^n + |\Sigma_{mi}| + |\Sigma_{ui}| + \underbrace{\widetilde{\boldsymbol{u}}_i^H \boldsymbol{C}_{mmi} \widetilde{\boldsymbol{u}}_i}_{\text{term I}} + \underbrace{\widetilde{\boldsymbol{u}}_i^H \boldsymbol{C}_{mui} \widetilde{\boldsymbol{v}}_i + \widetilde{\boldsymbol{v}}_i^H \boldsymbol{C}_{umi} \widetilde{\boldsymbol{u}}_i}_{\text{term II}} + \underbrace{\widetilde{\boldsymbol{v}}_i^H \boldsymbol{C}_{uui} \widetilde{\boldsymbol{v}}_i}_{\text{term III}} .
\end{aligned}
\tag{34}
$$

Concerning the second step, as $\widetilde{\boldsymbol{u}}_i$ and $\widetilde{\boldsymbol{v}}_i$ are independent, the conditional density of $\widetilde{\boldsymbol{v}}_i$ is in fact the marginal (Rice 1995) which is given by $\widetilde{\boldsymbol{v}}_i \sim \mathcal{N}_c\left(\boldsymbol{v}_i, \boldsymbol{\Sigma}_{ui}\right)$, however, as $\boldsymbol{v}_i$ is unknown and, in the conditional context, requires estimation conditional on knowing $\widetilde{\boldsymbol{u}}_i$ and $\boldsymbol{\vartheta}_k$, the conditional density of $\widetilde{\boldsymbol{v}}_i$ is expressed as $\widetilde{\boldsymbol{v}}_i \sim \mathcal{N}_c\left(\overline{\boldsymbol{v}}_i, \boldsymbol{\Sigma}_{ui}\right)$, where $\overline{\boldsymbol{v}}_i = \mathrm{E}\left[\boldsymbol{v}_i | \widetilde{\boldsymbol{u}}_i, \boldsymbol{\vartheta}_k\right]$. The EM algorithm requires the expectation of (34) over the probability space defined by this conditional PDF. Performing the expectation term by term, neglecting the terms that are constant with respect to $\boldsymbol{\vartheta}$, yields

$$
\begin{aligned}
\mathrm{E}\left[\text{term I} |\, \widetilde{\boldsymbol{u}}_i, \boldsymbol{\vartheta}_k\right] &= \widetilde{\boldsymbol{u}}_i^H \boldsymbol{C}_{mmi}\left(\boldsymbol{\vartheta}\right) \widetilde{\boldsymbol{u}}_i \\
\mathrm{E}\left[\text{term II} |\, \widetilde{\boldsymbol{u}}_i, \boldsymbol{\vartheta}_k\right] &= \widetilde{\boldsymbol{u}}_i^H \boldsymbol{C}_{mui}\left(\boldsymbol{\vartheta}\right) \overline{\boldsymbol{v}}_i + \overline{\boldsymbol{v}}_i^H \boldsymbol{C}_{mui}\left(\boldsymbol{\vartheta}\right) \widetilde{\boldsymbol{u}}_i \\
&= 2\mathbb{R}\mathrm{e}\left\{\widetilde{\boldsymbol{u}}_i^H \boldsymbol{C}_{mui}\left(\boldsymbol{\vartheta}\right) \overline{\boldsymbol{v}}_i\right\} \\
\mathrm{E}\left[\text{term III} |\, \widetilde{\boldsymbol{u}}_i, \boldsymbol{\vartheta}_k\right] &= \mathrm{tr}\left\{\boldsymbol{C}_{uui}\left(\boldsymbol{\vartheta}\right) \boldsymbol{\Sigma}_{ui}\right\} + \overline{\boldsymbol{v}}_i^H \boldsymbol{C}_{uui}\left(\boldsymbol{\vartheta}\right) \overline{\boldsymbol{v}}_i
\end{aligned}
$$

The integrations for the expectations of terms I and II are straightforward, but the expectation for term III is somewhat more complex, but it arises from a standard result in quadratic form theory for random variables (Mathai and Provost 1992).

Concerning the third step, it is required to determine an expression for $\overline{\boldsymbol{v}}_i$ dependent only on $\widetilde{\boldsymbol{u}}_i$ and $\boldsymbol{\vartheta}_k$. To do this, note that (26) and (21) imply that

$$
\boldsymbol{v}_i = -\boldsymbol{G}_{ui}^{+} \boldsymbol{G}_{mi} \boldsymbol{u}_i \tag{35}
$$

where $\boldsymbol{G}_{ui}^{+}$ is a Moore-Penrose pseudoinverse to $\boldsymbol{G}_{ui}$, which exists as $\boldsymbol{G}_{ui}$ is full column rank (Theorem 1). The equation (35) suggests the estimator $\overline{\boldsymbol{v}}_i = -\boldsymbol{G}_{ui}^{+}(\boldsymbol{\vartheta}_k)\boldsymbol{G}_{mi}(\boldsymbol{\vartheta}_k)\widetilde{\boldsymbol{u}}_i$ for which the expectation satisfies (35), meaning that it is an unbiased estimator. An appropriate expression for the Moore-Penrose inverse is $\boldsymbol{G}_{ui}^{+} = \boldsymbol{C}_{uui}^{-1}\boldsymbol{G}_{ui}^{H}\boldsymbol{\Lambda}_i^{-1}$ where $\boldsymbol{\Lambda}_i$ could be replaced by any nonsingular matrix of the correct size, but $\boldsymbol{\Lambda}_i$ was selected as it relates to the form of the MLEs (32). Defining $Q_i\left(\widetilde{\boldsymbol{u}}_i, \boldsymbol{\vartheta}_k, \boldsymbol{\vartheta}\right) = -\mathrm{E}\left[\ln\left(f_i\left(\widetilde{\boldsymbol{u}}_i, \widetilde{\boldsymbol{v}}_i | \widehat{\boldsymbol{u}}_i, \widehat{\boldsymbol{v}}_i, \boldsymbol{\vartheta}\right)\right) | \widetilde{\boldsymbol{u}}_i, \boldsymbol{\vartheta}_k\right]$, with the substitution for $\overline{\boldsymbol{v}}_i$ as outlined above and combining all the terms $i = 1, \ldots, M$, and using (23) leads to the expression (27).

APPENDIX III. MANUSCRIPT TABLES

TABLE 1. Categorization of nodes $\mathcal{N}$ for the network $(\mathcal{G}(\mathcal{N},\Lambda),\mathcal{P})$ into disjoint subsets based on whether the state variables of pressure ψ and flow θ are known, measured or unknown/unmeasured. Note that $\mathcal{A}_1 \cup \mathcal{A}_2 \cup \mathcal{A}_3 \cup \mathcal{A}_4 \cup \mathcal{B}_1 \cup \mathcal{B}_2 \cup \mathcal{C}_1 \cup \mathcal{C}_2 = \mathcal{N}$.

Nodal set	Nodal state classification		
	Known	Measured	Unknown/unmeasured
$\mathcal{A}_1$	-	ψ, θ	-
$\mathcal{A}_2$	-	ψ	θ
$\mathcal{A}_3$	-	θ	ψ
$\mathcal{A}_4$	-	-	ψ, θ
$\mathcal{B}_1$	θ	ψ	–
$\mathcal{B}_2$	θ	-	ψ
$\mathcal{C}_1$	ψ	θ	-
$\mathcal{C}_2$	ψ	-	θ

TABLE 2. The nodal partitioning for scenarios 1 and 2 for the 11-pipe network in Figure 1(b). The inclusion of $7 \in \mathcal{A}_4$ means that Scenario 1 correctly assumes that $\theta_7(t) \neq 0$, where as $7 \in \mathcal{B}_2$ for Scenario 2 incorrectly assumes that $\theta_7(t) = 0$.

Nodal set	Node sets for each case	
	Scenario 1	Scenario 2
$\mathcal{A}_1$	$\{4\}$	$\{4\}$
$\mathcal{A}_2$	$\emptyset$	$\emptyset$
$\mathcal{A}_3$	$\emptyset$	$\emptyset$
$\mathcal{A}_4$	$\{7\}$	$\emptyset$
$\mathcal{B}_1$	$\{2,6,3\}$	$\{2,6,3\}$
$\mathcal{B}_2$	$\{5\}$	$\{5,7\}$
$\mathcal{C}_1$	$\emptyset$	$\emptyset$
$\mathcal{C}_2$	$\{1\}$	$\{1\}$

TABLE 3. Sample statistics of parameter estimate relative errors for case study 1. The estimate sample statistics (median and the interquartile range, IQR) are based on 10 trials, where the presented relative errors of the estimates are averaged over all 11 pipes.

Parameter	Estimate sample statistic	Relative error statistics (%) (averaged over all 11 pipes)	
		Scenario 1	Scenario 2
r_o	median	30.68	307.26
	IQR	9.30	0.96
Γ_o	median	0.01	0.03
	IQR	0.01	0.00
Z_{co}	median	7.94	11.87
	IQR	3.46	1.50

TABLE 4. Sample statistics of parameter estimate relative errors for case study 2. The estimate sample statistics (median and the interquartile range, IQR) are based on 10 trials, where the presented relative errors of the estimates are averaged over all 11 pipes.

Parameter	Estimate sample statistic	Relative error statistics (%) (averaged over all 11 pipes)	
		Scenario 1	Scenario 2
wavespeed, c_o	median	0.313	0.411
	IQR	0.954	0.822
diameter, D	median	2.933	1.963
	IQR	0.392	0.377
length, l	median	0.311	0.415
	IQR	0.954	0.822
$\log_{10}\left(\frac{\epsilon}{D}\right)$	median	17.26	28.37
	IQR	35.51	15.21
$\log_{10}\mathbb{R}_e$	median	3.627	7.722
	IQR	4.970	1.709

APPENDIX IV. MANUSCRIPT ALGORITHMS

Algorithm 1 Computation of EM Parameter Estimate

Require: Network topology $\mathcal{G}(\mathcal{N}, \Lambda)$, frequency-domain data $\widetilde{\mathcal{U}}$ and corresponding frequencies $\boldsymbol{\omega} = \{\omega_1, \ldots, \omega_M\}$, covariances $\boldsymbol{\Sigma}_m$ and $\boldsymbol{\Sigma}_u$, parameter range $\boldsymbol{\Upsilon}$, initial parameter estimate $\boldsymbol{\vartheta}_0$;

Given $\mathcal{G}(\mathcal{N}, \Lambda)$, construct nodal sets $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ as defined in Table 1.
Reorder nodes in $\mathcal{N}$ from nodal sets as in (10);
$k = 0$; loop = `True`;
while loop **do**
    Compute updated parameter estimate as the maximiser of the expected conditional log likelihood
    $\boldsymbol{\vartheta}_{k+1} = \arg\max_{\boldsymbol{\vartheta}\in\boldsymbol{\Upsilon}} \left\{ \texttt{ExpCondLogLikelihood}(\boldsymbol{\vartheta}; \mathcal{G}(\mathcal{N}, \Lambda), \widetilde{\mathcal{U}}, \boldsymbol{\omega}, \boldsymbol{\Sigma}_m, \boldsymbol{\Sigma}_u, \boldsymbol{\vartheta}_k) \right\}$;
    if Termination criteria is satisfied **then**
        loop = `False`;
    end if
    $k = k + 1$;
end while
return Parameter estimate $\boldsymbol{\vartheta}_k$.

Algorithm 2 Computation of `ExpCondLogLikelihood` function from Algorithm 1

Require: New parameter estimate $\boldsymbol{\vartheta}$, Network topology $\mathcal{G}(\mathcal{N}, \Lambda)$, frequency-domain data $\widetilde{\mathcal{U}}$ and corresponding frequencies $\boldsymbol{\omega} = \{\omega_1, \ldots, \omega_M\}$, covariances $\boldsymbol{\Sigma}_m$ and $\boldsymbol{\Sigma}_u$, old parameter estimate $\boldsymbol{\vartheta}_k$;

Q = 0;
for $i = 1$ to M **do**
    for $\boldsymbol{\varphi} = \boldsymbol{\vartheta}$ and $\boldsymbol{\vartheta}_k$ **do**
        Compute admittance matrix $\boldsymbol{Y} = \boldsymbol{Y}(\mathrm{i}\omega_i, \boldsymbol{\varphi})$ from (15);
        Construct submatrices $\boldsymbol{Y}_{m1}, \boldsymbol{Y}_{m2}, \boldsymbol{Y}_{m3}, \boldsymbol{Y}_{u1}, \boldsymbol{Y}_{u2}$, and $\boldsymbol{Y}_{u3}$ as in (19) from submatrices of $\boldsymbol{Y}$ as in (16);
        Construct $\boldsymbol{G}_m$ and $\boldsymbol{G}_u$ as in (18);
        Compute $\boldsymbol{\Lambda}_i(\boldsymbol{\varphi})$ as in (30);
        Compute $\boldsymbol{C}_{mmi}(\boldsymbol{\varphi}), \boldsymbol{C}_{mui}(\boldsymbol{\varphi}), \boldsymbol{C}_{umi}(\boldsymbol{\varphi})$, and $\boldsymbol{C}_{uui}(\boldsymbol{\varphi})$ as in (29);
    end for
    Compute Q_i as in (28);
    $Q = Q + Q_i$
end for
return Expected conditional log likelihood value $-Q$.

APPENDIX V. MANUSCRIPT FIGURES

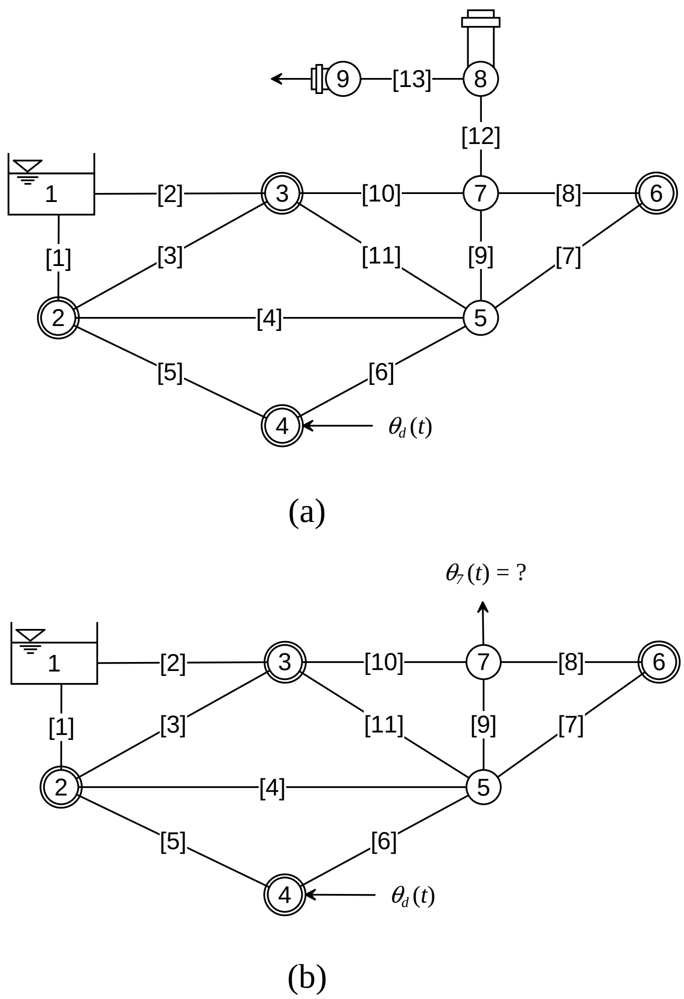

FIG. 1. The extended 11-pipe network. The double rings around the nodes indicate the locations of pressure measurements. Subfigure (a) represents the actual true network which is the 11-pipe network from with an additional branch from node 7 consisting of two pipes with a capacitor at node 8 and an emitter at node 9. Subfigure (b) represents the known configuration of the network involving an unknown nodal flow at node 7, as the existence of the connection is known but the form of the subnetwork outside this node is unknown.

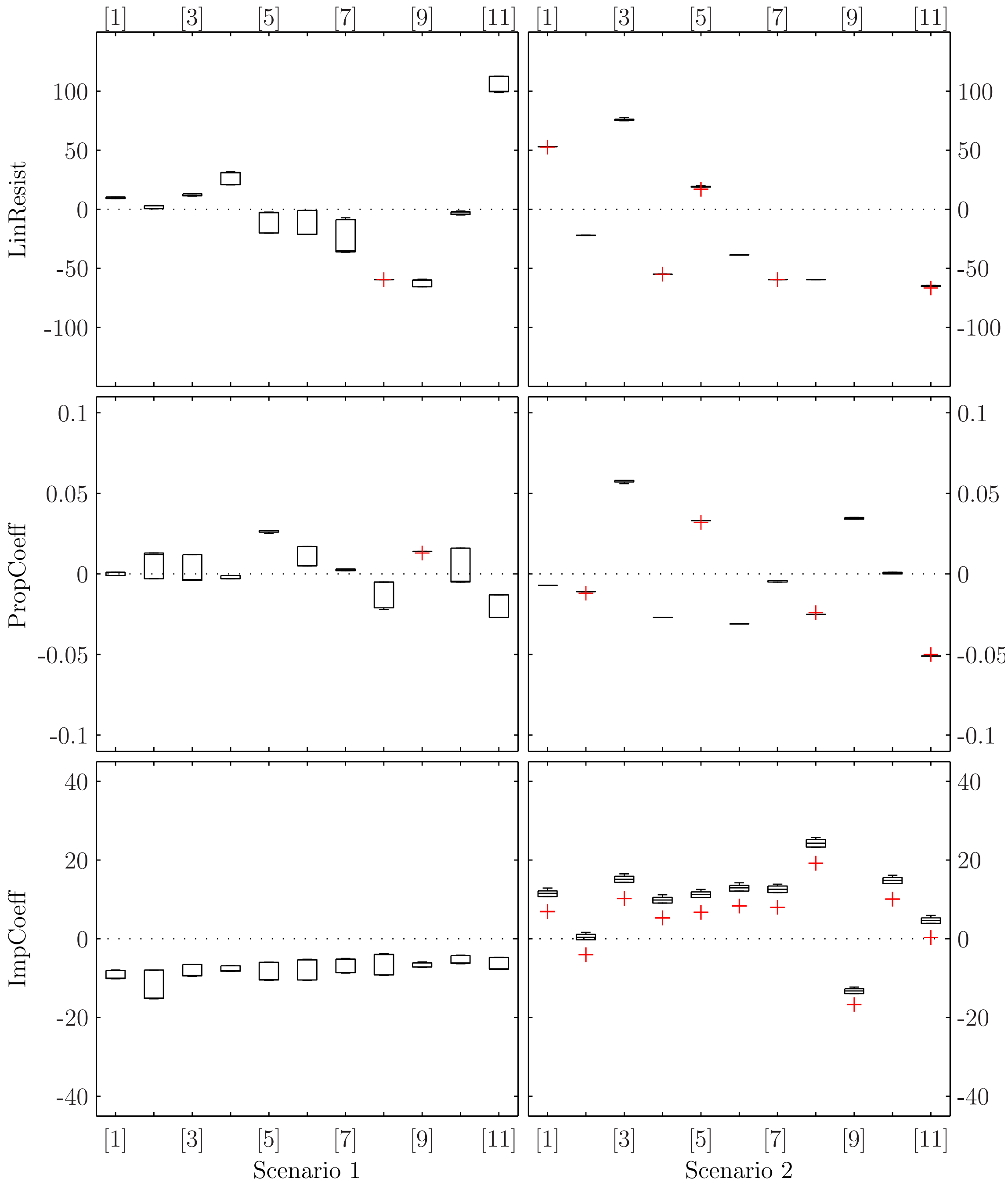

FIG. 2. Box plots of the relative errors (%) of the parameter estimates from case study 1 (TSF model). Within each subfigure, the vertical axis gives the relative error, and the horizontal axis indicates the pipe number (i.e. each box and whisker set is associated with a pipe parameter estimate). Each subfigure row is associated with a particular parameter (indicated to the left of the subfigure matrix), and each subfigure column with particular scenarios (indicated on the bottom of the subfigure matrix). For each case, 10 independent trials were performed. The + indicate outliers.

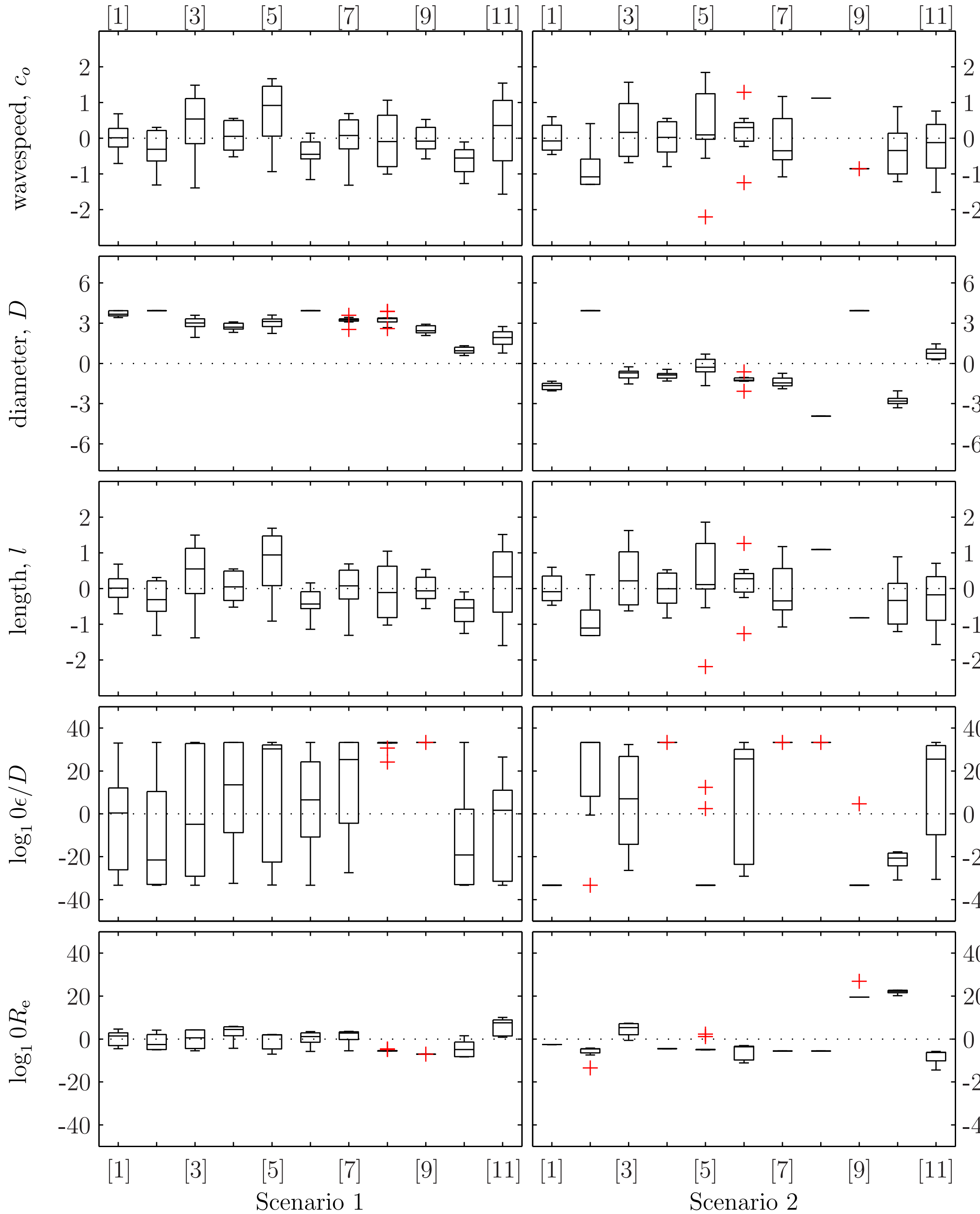

FIG. 3. Box plots of the relative errors (%) of the parameter estimates from case study 2 (TUF model). Within each subfigure, the vertical axis gives the relative error, and the horizontal axis indicates the pipe number (i.e. each box and whisker set is associated with a pipe parameter estimate). Each subfigure row is associated with a particular parameter (indicated to the left of the subfigure matrix), and each subfigure column with particular scenarios (indicated on the bottom of the subfigure matrix). For each case, 10 independent trials were performed. The + indicate outliers.