



# Data-driven discovery of PDEs in complex datasets

Jens Berg, Kaj Nyström\*

Department of Mathematics, Uppsala University, SE-751 05 Uppsala, Sweden



## ARTICLE INFO

### Article history:

Received 31 August 2018

Received in revised form 8 January 2019

Accepted 16 January 2019

Available online 25 February 2019

### Keywords:

Machine learning

Deep learning

Partial differential equations

Neural networks

## ABSTRACT

Many processes in science and engineering can be described by partial differential equations (PDEs). Traditionally, PDEs are derived by considering first principles of physics to derive the relations between the involved physical quantities of interest. A different approach is to measure the quantities of interest and use deep learning to reverse engineer the PDEs which are describing the physical process.

In this paper we use machine learning, and deep learning in particular, to discover PDEs hidden in complex data sets from measurement data. We include examples of data from a known model problem, and real data from weather station measurements. We show how necessary transformations of the input data amounts to coordinate transformations in the discovered PDE, and we elaborate on feature and model selection. It is shown that the dynamics of a non-linear, second order PDE can be accurately described by an ordinary differential equation which is automatically discovered by our deep learning algorithm. Even more interestingly, we show that similar results apply in the context of more complex simulations of the Swedish temperature distribution.

© 2019 Elsevier Inc. All rights reserved.

## 1. Introduction

Modern technology has made high-quality data available in abundance. It is estimated that more than 2.5 quintillion bytes of data is generated every day and that 90% of all data were generated in the last two years alone [2]. The amount of user generated data on social media and data generated through smart sensors in the Internet of things will likely contribute to an even faster increase. A major problem of scientific and industrial interest is how to transform the data into a predictive model which can give insights on the data generating process.

The data generating process in the natural sciences is often described in terms of differential equations. There is a vast amount of literature spanning over decades available for the identification of dynamical systems where the quantities of interest are measured as a function of time or some other dependent variable. See for example [41,38,20,24,9,37,17]. The identification of time-dependent partial differential equations (PDEs) through data analysis is an emerging and exciting field of research which is not as explored as dynamical systems. The research has been made available through the recent progress in machine learning algorithms and their efficient implementation in open source software.

PDEs are traditionally derived by considering first physical principles. For example the Navier-Stokes equations in fluid dynamics are derived by considering the conservation of mass, momentum, and energy for a control volume in a fluid. There are, however, many situations where derivations by first principles are intractable or even impossible as they become

\* Corresponding author.

E-mail addresses: [jens.berg@math.uu.se](mailto:jens.berg@math.uu.se) (J. Berg), [kaj.nystrom@math.uu.se](mailto:kaj.nystrom@math.uu.se) (K. Nyström).

too complicated or the governing physical laws are unknown. In such situations there are typically several geostationary points where changes of quantities of interest are measured over time. Datasets consisting of such spatio-temporal data is the interest of this paper and we aim to develop methods which can automatically identify a PDE which is generating the dataset.

The purpose of identifying a partial differential equation is to be able to make predictions of physical quantities outside the range of the measurement data. A function which is approximating the measured data can be accurately interpolated by a variety of methods, for example finite differences [11], polynomial interpolation [33] [10], finite elements, spectral methods, radial basis functions, or neural networks [30,31]. The interpolation is in general, however, only accurate within the range of the data. Differential operators, on the other hand, can be used anywhere as long as the corresponding differential equation can be solved.

The emerging field of data-driven discovery of PDEs can be split into three approaches: (1) Sparse regression, (2) Gaussian processes, and (3) Artificial neural networks. Sparse regression is based on a library of candidate terms and sparse model selection to select the most important terms [36,33,35]. Identification using Gaussian processes works by placing a Gaussian process prior on the unknown coefficients of the PDE and infer them by using maximum likelihood estimation [29,32,28]. Artificial neural networks can be used as sparse regression models, act as priors on unknown coefficients, or completely determine a general differential operator [4,27,31].

In this paper we will focus on deep neural networks to extend and complement previous work mentioned in the above references. There are two main contribution in this paper. The first is that we use a unified neural network approach for both sparse regression and the identification of general differential operators. The second is that we include complex datasets where necessary transformations of the input data manifest as coordinate transformations which yield metric coefficients in the identified PDE.

## 2. Method

We are working under the assumption that we have an unordered dataset consisting of space-time coordinates and function values where the governing equation is unknown. The goal is to identify a PDE which approximately has the function values as the solution in the space-time points. The first step is to fit a function to the data which can be used to compute the derivatives with respect to the space-time coordinates. This is a separate preprocessing step and any method can be used. The most recent work has been focused on polynomial interpolation or neural networks due to their independence of structured data and insensitivity to noise.

The identified PDE depends highly on the quality of the approximating function and a comparative study of various approximation methods would be valuable and is the topic of future research. We will use deep neural networks as approximating functions. Deep neural networks are universal smooth function approximators [14,15,6] and their derivatives are analytically available through backpropagation [34,18] or automatic differentiation [41] in open source software such as TensorFlow [3] or PyTorch [25]. Since the neural network is only used to compute derivatives of the interpolated solution in the range of the training data, without the need for generalization, overfitting is not a major issue in the first step. On the contrary, overfitting could be useful for obtaining a maximally accurate interpolation of the training data. We stress that accurate interpolation of the measurement data is necessary to compute accurate derivatives which are used when identifying the differential operator. The identified differential operator will not be more accurate than the interpolation of the measurement data.

We assume that our data consists of the triplets  $t$ ,  $\mathbf{x} = [x_1, x_2, \dots, x_N]$ , and  $\mathbf{u} = [u_1, u_2, \dots, u_M]$  which is describing a vector valued mapping  $\mathbf{u} : \mathbb{R}^{N+1} \rightarrow \mathbb{R}^M$ , where  $t$  denotes the time variable,  $x_1, \dots, x_N$  the space variables, and  $u_1, \dots, u_M$  the function values. In the first step we approximate the function  $u$  by a deep neural network  $\hat{u} = \hat{u}(\mathbf{x}, t; \mathbf{p})$  where  $\mathbf{p}$  denotes the vector of coefficients in the network. We will usually drop explicit parameter dependence, unless required, to ease the notation. We will use the hyperbolic tangent as activation function and solve the regularized minimization problem for the coefficients,

$$\mathbf{p}^* = \min_{\mathbf{p}} \frac{1}{2} \|u(\mathbf{x}, t) - \hat{u}(\mathbf{x}, t; \mathbf{p})\|^2 + \frac{\alpha_p}{2} \|\mathbf{p}\|^2, \quad (2.1)$$

by using the BFGS [8] or L-BFGS [19] methods for small and large scale problems, respectively. When solving the minimization problem (2.1), we do not distinguish between the time and space coordinates. Different datasets require different neural networks designs and it would be interesting to try neural networks which are tailored for time-series prediction, for example recurrent neural networks, in this context. Such a study is, however, beyond the scope of this paper.

In the second step we seek a parameterized function  $\hat{L} = \hat{L}(\hat{u}, \partial \hat{u}, \dots, \partial^m \hat{u}; \mathbf{q})$ , where the notation  $\partial^j \hat{u}$  means all partial derivatives of  $\hat{u}$  with respect to  $x_1, \dots, x_n$  up to order  $m$  such that

$$\hat{u}_t = \hat{L}(\hat{u}, \partial \hat{u}, \dots, \partial^m \hat{u}). \quad (2.2)$$

$\hat{L}$  is then the approximation of the, yet unknown, differential operator in the governing PDE. The restriction to first order time derivatives is without loss of generality as we can compute derivatives of any order from the neural network approximation  $\hat{u}$ . Note that we in (2.2) for simplicity assume that there is no space-time dependence on the parameters in the differential operator. This is not an essential restriction since we could equally well have the operator on the form

$$\hat{u}_t = \hat{L}(x, t, \hat{u}, \partial \hat{u}, \dots, \partial^m \hat{u}) \quad (2.3)$$

where space-time dependence is included. However, the resulting operator is rarely suitable for human interpretation and the explicit space-time parameter dependence can in general not be discerned. Neural networks are on the other hand suitable for coefficient inverse problems where the form of the operator is known but the space-time dependent parameters are unknown. See for example [4]. The space-time dependent form (2.3) is more sensitive to overfitting and the form (2.2) should primarily be considered unless there is a reason not to.

Depending on the choice of parametrization of  $\hat{L}$  it is possible to discover a wide range of PDEs and encompass the methods described in [33,5,35,36,27,31] in a single framework. The framework we have chosen here is to represent  $\hat{L}$  by a feedforward neural network and to find  $\hat{L}$  by gradient based optimization. We recover the sparse regression method by having a neural network without hidden layers with candidate terms as input features, in which case the neural network reduces to a linear model. We recover classical PDEs, which are polynomial in  $\hat{u}$  and its partial derivatives, by computing all partial derivatives up to some order  $m$ , all non-linear combinations up to some order  $k$ , and having them as input features to a linear model. There are

$$\mathcal{M} = M \left( 1 + \sum_{i=1}^m \binom{i+N-1}{N-1} \right)$$

partial derivative terms up to order  $m$  and

$$\mathcal{K} = \sum_{i=1}^k \binom{i+\mathcal{M}-1}{\mathcal{M}-1}$$

non-linear combinations up to order  $k$ . For example, the time-dependent compressible Navier-Stokes equations in 3D have  $N = 3$  space variables,  $M = 5$  unknowns, non-linear terms up to order  $k = 2$ , and partial derivatives to up to order  $m = 2$ . This gives a total of  $\mathcal{M} = 50$  partial derivative terms and  $\mathcal{K} = 1325$  possible input features. While the number of input features grows combinatorially with the number of partial derivatives and non-linear order, modern day machine learning with neural networks casually deal with input features in the order of million or even billions. Even the most basic standard example of hand written digit recognition using the MNIST dataset has  $28 \times 28 = 784$  input features – the number of pixels of each image in the dataset. Finally, we can let  $\hat{L}$  be given by a neural network of arbitrary complexity with the  $\mathcal{M}$  partial derivative terms as input features to get an arbitrarily complex differential operator.

There is always a trade-off between model complexity and interpretability. A linear model with candidate terms as input features provides a simple model which can be read, analyzed, and understood. It does, however, require some physical understanding of the data generating process to ensure that the set of input features is sufficient. A general neural network model is on the other extreme. It can approximate an arbitrary complex differential operator but the resulting operator can neither be read nor understood. A linear model with polynomial input features is somewhere in between. Sparse regression with L1 regularization will remove some insignificant terms but some manual post cleaning will probably be required to get a interpretable model. In all cases, the model is unlikely to produce a well-posed PDE in the sense of Hadamard [13].

As the true differential operator  $L$  is not known and we have no training data for it, the goal is to find a set of parameters  $\mathbf{q}^*$  such that the residual of the approximate PDE is minimized,

$$\mathbf{q}^* = \min_{\mathbf{q}} \frac{1}{2} \|\hat{u}_t - \hat{L}(\hat{u}, \partial \hat{u}, \dots, \partial^m \hat{u}; \mathbf{q})\|^2 + \frac{\alpha_q}{2} \|\mathbf{q}\|_1^2. \quad (2.4)$$

We typically add regularization in the  $L^1$ -norm to favor sparsity in the resulting PDE model. The optimization problems (2.1) and (2.4) are very different from an optimization perspective. The former is a highly non-convex optimization problem over a large number of parameters and a limited amount of data. The latter is, in the linear model case, a convex optimization problem over a small number of parameters and a large amount of data. In the 3D Navier-Stokes example above, let us assume that we have sampled the solution 100 times on a  $32 \times 32 \times 32$  grid. This gives us a dataset of size  $3276800 \times 4$  in the optimization of (2.1) and  $3276800 \times 1325$  in the optimization of (2.4). Data driven discovery of PDEs is thus suitable on heterogeneous systems where the optimization of (2.1) is performed on GPUs with many cores and limited memory while the optimization of (2.4) is performed on CPUs with few cores and large memory.

## 2.1. Feature scaling

It is well-known that machine learning algorithms perform poorly unless the input features are scaled correctly. In the previous work on data-driven discovery of PDEs, all data were generated by known PDEs on simple geometries which did not require any transformation of the input features. In real life applications, however, the domain of interest is in general neither simple nor close to the origin and the input features need to be transformed. The transformation then impacts the identified PDE as it is subjected to a coordinate transformation. Using a neural network to approximate the dataset as a separate preprocessing step usually follows a pipeline in which feature scaling is included, for example by preprocessing

using the `Pipeline` module from `scikit-learn` [26]. It is hence important to be aware of all feature scalings in the preprocessing step and that the exact same feature scaling is used in the identification of the PDE in the second step.

Feature scaling amounts to the invertible coordinate transformations

$$\begin{aligned}\tau &= \tau(t), \\ \xi &= \xi(\mathbf{x})\end{aligned}\tag{2.5}$$

where  $\tau, \xi = [\xi_1, \dots, \xi_N]$  are the new time and space coordinates, respectively. A common transformation is to shift and scale such that each input feature has zero mean and unit variance,

$$\begin{aligned}\tau &= \frac{t - \bar{t}}{\sigma(t)}, \\ \xi &= \frac{\mathbf{x} - \bar{\mathbf{x}}}{\sigma(\mathbf{x})},\end{aligned}\tag{2.6}$$

where  $\bar{t}, \bar{\mathbf{x}}$  and  $\sigma(\cdot)$  denotes the (componentwise) average and standard deviation of the input data, respectively, and the division is performed componentwise where needed.

As an example we can consider what happens to the discovery of the viscous Burger's equation under the transformation (2.5). Assume we are given a dataset generated by the viscous Burger's equation in 1D,

$$u_t + uu_x = \epsilon u_{xx},\tag{2.7}$$

to which we fit a neural network under the general coordinate transformation (2.5). By the chain rule we get

$$\begin{aligned}\frac{\partial u}{\partial t} &= \frac{\partial u}{\partial \tau} \frac{\partial \tau}{\partial t}, \\ \frac{\partial u}{\partial x} &= \frac{\partial u}{\partial \xi} \frac{\partial \xi}{\partial x}, \\ \frac{\partial^2 u}{\partial x^2} &= \frac{\partial^2 u}{\partial \xi^2} \left( \frac{\partial \xi}{\partial x} \right)^2 + \frac{\partial u}{\partial \xi} \frac{\partial^2 \xi}{\partial x^2}\end{aligned}$$

and hence the neural network is not an approximation to the solution of (2.7) but to the transformed equation

$$\frac{\partial \tau}{\partial t} u_\tau + \left( \frac{\partial \xi}{\partial x} u - \epsilon \frac{\partial^2 \xi}{\partial x^2} \right) u_\xi = \epsilon \left( \frac{\partial \xi}{\partial x} \right)^2 u_{\xi\xi}.$$

Under the linear transformation (2.6), the above equation reduces to

$$\frac{1}{\sigma(t)} u_\tau + \frac{1}{\sigma(x)} uu_\xi = \frac{\epsilon}{\sigma^2(x)} u_{\xi\xi}.\tag{2.8}$$

The situation becomes more complex in higher dimensions as in general we need to compute all total derivatives in the old coordinates when computing the partial derivatives in the new coordinates as

$$\begin{aligned}\frac{\partial u}{\partial x_1} &= \frac{\partial u}{\partial \xi_1} \frac{\partial \xi_1}{\partial x_1} + \dots + \frac{\partial u}{\partial \xi_N} \frac{\partial \xi_N}{\partial x_1}, \\ &\vdots \\ \frac{\partial u}{\partial x_N} &= \frac{\partial u}{\partial \xi_1} \frac{\partial \xi_1}{\partial x_N} + \dots + \frac{\partial u}{\partial \xi_N} \frac{\partial \xi_N}{\partial x_N}.\end{aligned}$$

We write the above expression in matrix form as

$$\begin{bmatrix} \frac{\partial u}{\partial x_1} \\ \vdots \\ \frac{\partial u}{\partial x_N} \end{bmatrix} = \begin{bmatrix} \frac{\partial \xi_1}{\partial x_1} & \dots & \frac{\partial \xi_N}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial \xi_1}{\partial x_N} & \dots & \frac{\partial \xi_N}{\partial x_N} \end{bmatrix} \begin{bmatrix} \frac{\partial u}{\partial \xi_1} \\ \vdots \\ \frac{\partial u}{\partial \xi_N} \end{bmatrix}$$

where the square matrix above is the Jacobian matrix,  $J$ , of the coordinate transformation. Since we are interested in the PDE in the physical coordinates, we need to transform back to the original coordinates by computing the inverse of the Jacobian,

$$J^{-1} = \begin{bmatrix} \frac{\partial x_1}{\partial \xi_1} & \cdots & \frac{\partial x_N}{\partial \xi_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_1}{\partial \xi_N} & \cdots & \frac{\partial x_N}{\partial \xi_N} \end{bmatrix}.$$

The transformation (2.6) is particularly useful in high dimensions as it is linear and acts only one coordinate direction at a time, independently of the other coordinates. This means that the Jacobian is reduced to the diagonal matrix

$$J = \begin{bmatrix} \frac{1}{\sigma(x_1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\sigma(x_N)} \end{bmatrix}$$

and that higher-order derivatives are easily computed since each derivative of  $u$  with respect to  $x_i$  only yields an additional factor of  $1/\sigma(x_i)$ . That is, we get

$$\begin{aligned} \frac{\partial u}{\partial x_i} &= \frac{1}{\sigma(x_i)} \frac{\partial u}{\partial \xi_i}, \\ \frac{\partial^2 u}{\partial x_i \partial x_j} &= \frac{1}{\sigma(x_i) \sigma(x_j)} \frac{\partial^2 u}{\partial \xi_i \partial \xi_j} \\ &\vdots \\ \frac{\partial^m u}{\partial x_i \cdots \partial x_j} &= \frac{1}{\sigma(x_i) \cdots \sigma(x_j)} \frac{\partial^m u}{\partial \xi_i \cdots \partial \xi_j} \end{aligned}$$

for the partial derivatives up to order  $m$ . Transforming the partial derivatives back to the original coordinates is reduced to multiplication by a scalar which avoids the numerically unstable and computationally expensive inversion of the Jacobian matrix.

### 3. Examples

There are plenty of examples in previous papers which show impressive results in the accuracy of the identified PDE despite both sparse and noisy data [33,5,35,36,27,31]. These results are all based on known PDEs on simple geometries. We will show a few examples on what happens to the identified PDE under coordinate transformations, and some potential applications in weather/climate modeling where the governing equations are unknown.

#### 3.1. The viscous Burger's equation in 1D

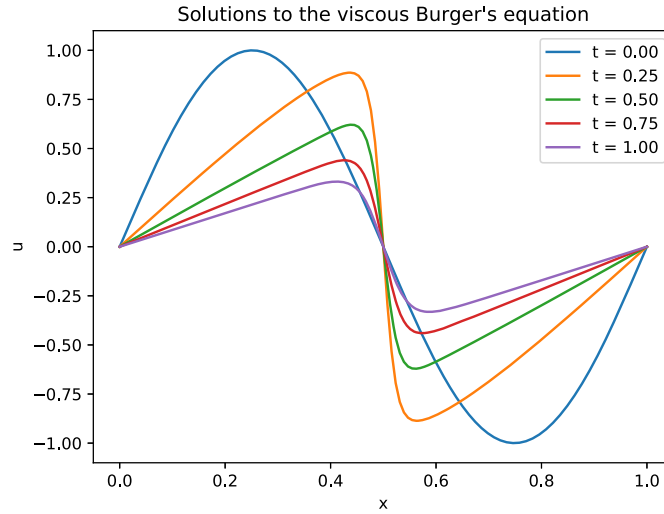
We consider the viscous Burger's equation for  $(x, t) \in [0, 1] \times [0, 1]$  here given by

$$\begin{aligned} u_t + uu_x &= 10^{-2} u_{xx}, \\ u(0, t) &= 0, \\ u(1, t) &= 0, \\ u(x, 0) &= \sin(2\pi x). \end{aligned} \tag{3.1}$$

The solution to (3.1) is well-known and forms a decaying stationary viscous shock after a finite time, see Fig. 1.

The solution of (3.1) was computed with the finite element method using 128 second-order elements in space and 1000 steps using the backward Euler method in time.

To reconstruct the differential operator in (3.1), we sample the solution in all interior degrees of freedom at each non-zero time step to get a dataset of the form  $(t, x, u)$  consisting of a total of 255000 entries. The first step is to fit a neural network to the dataset which allows us to compute the necessary derivatives. This is a separate preprocessing step in which we use a feedforward neural network with 5 hidden layers and 10 neurons in each layer with the hyperbolic tangent activation function. The network is trained using the BFGS method from SciPy's `scipy.optimize` module with default parameters [16]. For this model problem we consider three different parametrizations of  $\hat{L}$  without regularization or scaling: 1) A linear model with the library terms  $uu_x$  and  $u_{xx}$  as input features, 2) A linear model with up to second order derivative and non-linear terms as input features, and 3) A single layer feedforward neural network with 5 neurons in the hidden



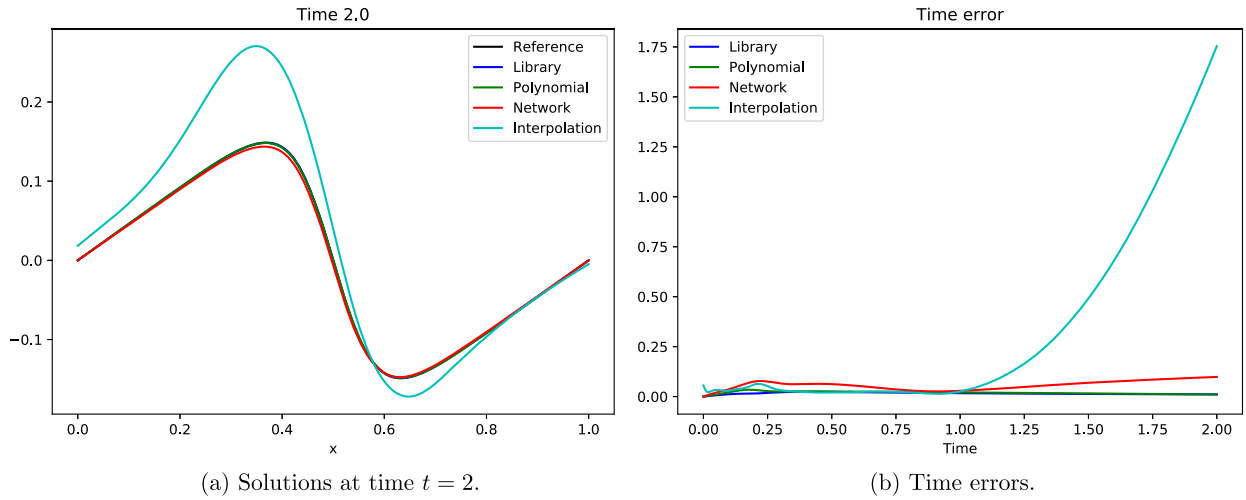
**Fig. 1.** The solution of the viscous Burger's equation forming a stationary viscous shock. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

layer with the hyperbolic tangent activation and up to second order derivative terms as input features. When minimizing (2.4) with the different parametrizations we discover the following PDEs:

- 1)  $u_t = -9.9792 \times 10^{-1} * uu_x + 9.9596 \times 10^{-3} * u_{xx}$
- 2)  $u_t = -9.9718 \times 10^{-1} * uu_x + 1.0134 \times 10^{-2} * u_{xx}$   
 $- 4.2757 \times 10^{-8} * (u_{xx})^2 + 1.0156 \times 10^{-5} * u_x u_{xx}$   
 $- 8.3758 \times 10^{-5} * uu_{xx} + 2.8494 \times 10^{-6} * (u_x)^2$   
 $+ 1.4114 \times 10^{-4} * u_x - 4.4878 \times 10^{-3} * (u)^2 + 2.2429 \times 10^{-3} * u$
- 3)  $u_t = 1.6596 \times 10^2 * \tanh(-6.3509 \times 10^{-5} * u_{xx} - 6.3461 \times 10^{-3} * u_x$   
 $- 7.9442 \times 10^{-1} * u + 2.4650)$   
 $- 2.0935 \times 10^2 * \tanh(-6.2849 \times 10^{-5} * u_{xx} - 7.6160 \times 10^{-3} * u_x$   
 $+ 3.3920 \times 10^{-1} * u + 1.2439)$   
 $- 3.5269 \times 10^2 * \tanh(-2.2973 \times 10^{-6} * u_{xx} - 3.7830 \times 10^{-3} * u_x$   
 $- 2.0181 \times 10^{-1} * u - 4.8004 \times 10^{-1})$   
 $+ 2.7530 \times 10^2 * \tanh(9.9351 \times 10^{-6} * u_{xx} - 4.6315 \times 10^{-3} * u_x$   
 $- 3.0447 \times 10^{-1} * u + 4.2278 \times 10^{-1})$   
 $+ 4.5559 \times 10^2 * \tanh(3.2915 \times 10^{-5} * u_{xx} - 3.7305 \times 10^{-3} * u_x$   
 $+ 3.3400 \times 10^{-1} * u - 1.3281)$   
 $+ 1.4213 \times 10^2$

It is clear that the different models have different trade-offs. The first model is similar in appearance to the true PDE, but it is required that we know the form of the PDE a priori. The second model has small coefficients for the spurious terms and close to the true values for the true terms. The third model is general and of limited use for human interpretation. However, many PDE solvers offer automatic discretization of symbolic expressions and the output of the general model can be used as input to a software such as Comsol Multiphysics [1], or physics informed neural networks [30], or finite differences.

In Fig. 2, we have solved the PDEs (1)–(3) above using finite differences on summation-by-parts (SBP) form. SBP finite differences are suitable for automatic discretization of general PDEs due to their operator form where derivatives are computed by pure matrix-vector multiplications. See for example [39,7] and references therein. In this case, we used second order accurate SBP operators for the space discretization with strong boundary condition implementation on a mesh con-



**Fig. 2.** The solutions and errors of the Burger's equation using some different discovered operators including the interpolated neural network solution.

sisting of 1001 grid points. We performed the time integration using the classical 4th order Runge-Kutta method. We used a sufficient number of points in space to avoid dispersion errors which otherwise occur with central finite differences without artificial dissipation for non-linear equations, and a sufficiently small time step for the solution to remain stable. We can clearly see the benefit of using the discovered operators compared to the interpolated solution. The interpolated solution fails to approximate the solution beyond the range of the training data. The operators, however, are able to give accurate solutions for any range. We can see a slight increase of the error for the general neural network operator due to some overfitting.

**Remark 3.1.** In this work, we assumed the general operator (2.2). It is, however, possible to enforce the operator to be in conservative form

$$\hat{u}_t = -\nabla \cdot \hat{F}(\hat{u}),$$

where instead  $\hat{F}$  is the flux of  $\hat{u}$ . Such a flux formulation can then be almost trivially implemented in a finite element framework such as FEniCS [21].

To see the effect of a feature scaling we consider the simple library model under the standard shift and scale transformation (2.6). For this particular dataset we have

$$\begin{aligned}\sigma^2(t) &= 0.0833326, & \sigma^2(x) &= 0.08268167, \\ \bar{t} &= 0.50050196, & \bar{x} &= 0.49999807,\end{aligned}$$

and the identified PDE in transformed space becomes

$$u_\tau = -1.0010uu_\xi + 3.4815 \times 10^{-2}u_{\xi\xi}$$

which corroborates (2.8) rewritten as

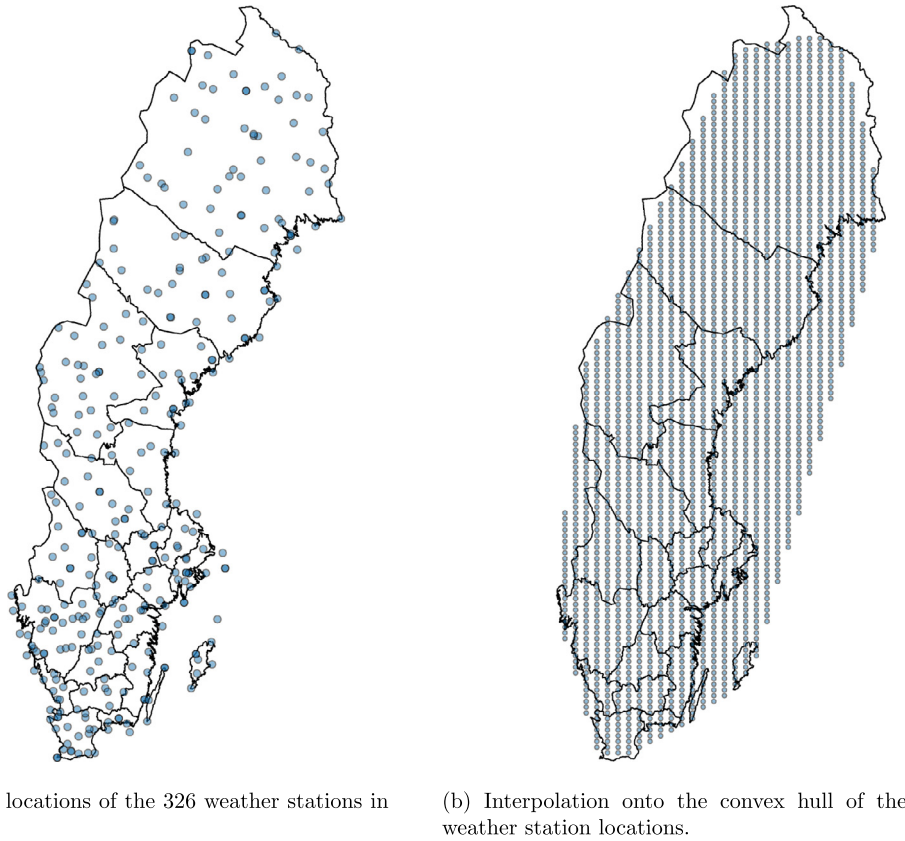
$$u_\tau = -\frac{\sigma(t)}{\sigma(x)}uu_\xi + \frac{\sigma(t)}{\sigma^2(x)} \times 10^{-2}u_{\xi\xi}.$$

To get the PDE in the physical coordinates it is hence required that we invert the coordinate transformation and compute the derivatives in the physical space as

$$\begin{aligned}\frac{\partial u}{\partial \tau} &= \frac{\partial u}{\partial t} \frac{\partial t}{\partial \tau} = \sigma(t) \frac{\partial u}{\partial t} \\ \frac{\partial u}{\partial \xi} &= \frac{\partial u}{\partial x} \frac{\partial x}{\partial \xi} = \sigma(x) \frac{\partial u}{\partial x} \\ \frac{\partial^2 u}{\partial \xi^2} &= \frac{\partial^2 u}{\partial x^2} \left( \frac{\partial x}{\partial \xi} \right)^2 = \sigma^2(x) \frac{\partial^2 u}{\partial x^2}.\end{aligned}$$

First after transforming back to the physical space do we recover the desired PDE





**Fig. 3.** Physical and interpolated locations of the geostationary locations.

$$u_t + \frac{\sigma(x)}{\sigma(t)} u u_x = \frac{\sigma^2(x)}{\sigma(t)} u_{xx},$$

and in this particular case we get

$$u_t + 0.99708 u u_x = 0.99717 \times 10^{-2} u_{xx}.$$

For this model problem, coordinate transformations are not necessary as we are working on the simple domain  $(x, t) \in [0, 1] \times [0, 1]$  which is in the range where machine learning algorithms performs well.

### 3.2. Temperature distribution in 2D

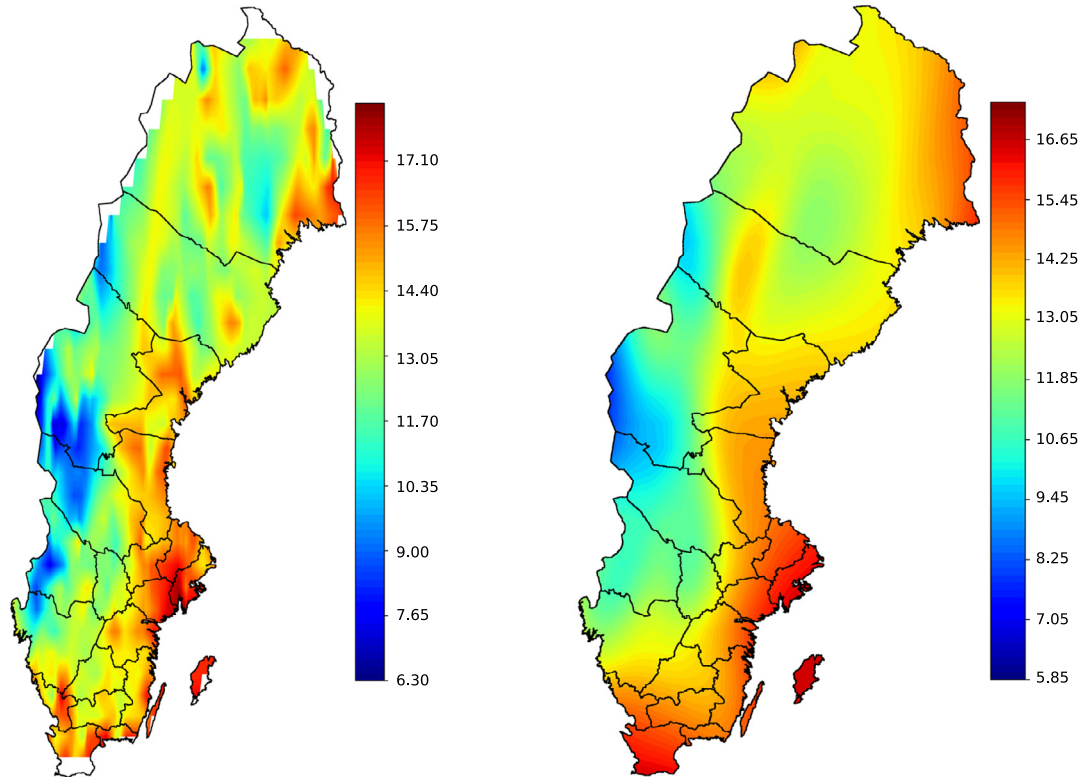
The focus in this section is on potential applications of the method to real measurement data for weather/climate modeling. The outlined method should be seen as a starting point for further research.

A natural application of the method is where several geostationary sensors are recording measurements over time, for example weather stations which measure quantities such as temperature, pressure, humidity, and so on on a regular basis. The Swedish Meteorological and Hydrological Institute<sup>1</sup> is offering a REST API where meteorological data can be downloaded for all 326 measurement stations in Sweden. Each station is recording data at time intervals ranging from every hour to every 12 hours, and the locations are given in latitude/longitude coordinates in the range  $[10.96, 55.34] \times [24.17, 69.05]$  which is outside the range where machine learning algorithms perform well. We downloaded the data and made a dataset consisting of the temperature for the first week in July 2016 to see if we can find a PDE which is describing the temperature distribution (Fig. 3).

The dataset contains irregular measurements in a complicated geometry where coordinate transformations are inevitable. The dataset is imbalanced since there are too many points in time compared to the number of points in space, and we were in fact unable to find a neural network which could accurately interpolate the original dataset. In this artificial example, we remedy this by performing a linear interpolation in space and time onto the convex hull of a regular grid with 168 time points, 32 latitude points, and 128 longitude points, see Fig. 4 (where all spatial data points have been transformed

<sup>1</sup> <http://www.smhi.se>.





(a) Linear interpolation temperature snapshot.

(b) Neural network approximated temperature snapshot.

**Fig. 4.** The linear interpolation and neural network approximated temperature snapshots. The neural network has 5 layers with 20 neurons each.

by the Mercator projection for visualization only). The final interpolated dataset contains 688129 data points on a regular grid. Finally, we approximate the dataset with a neural network with 5 hidden layers with 20 neurons in each layer using the L-BFGS optimization method. We tried many different networks and this, surprisingly small network, had the best generalization accuracy when evaluated on different test sets obtained by different interpolations. Larger networks had problems with overfitting and adding dropout and regularization caused the L-BFGS algorithm to perform poorly.

Note that since the neural network is globally defined we can plot the temperature in the whole domain and not just on the convex hull of the data points. In this case, the governing PDE is unknown and we will elaborate on results and conclusions in section 5.

**Remark 3.2.** The construction of artificial data is a research topic on its own and we will not elaborate further on the quality of the linear interpolation for this example. The linear interpolation is certainly not accurate and will give rise to non-physical relations which will be captured by the neural network interpolation and transferred to the discovered operator. The example is meant to illustrate that the same observations and conclusions as in the model problem case of the viscous Burger's equation transfers to more complicated datasets, as will be seen in Section 5.

#### 4. Feature selection

To elaborate on feature selection we return to Section 3.1 and the polynomial PDE model for the viscous Burger's equation which has a decent trade-off between complexity and interpretability. By adding L1 regularization to the polynomial PDE model with  $\alpha_q = 10^{-2}$  in (2.4), the spurious terms are further reduced to

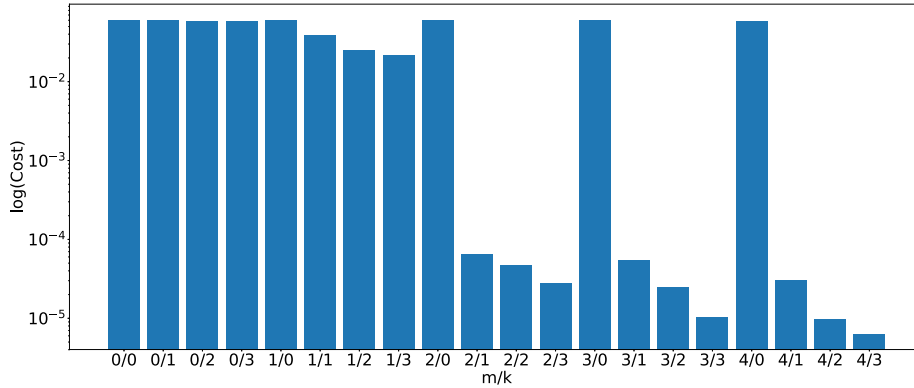
$$\begin{aligned}
 u_t = & -9.9473 \times 10^{-1} * uu_x + 1.0105 \times 10^{-2} * u_{xx} \\
 & - 1.2630 \times 10^{-9} * (u_{xx})^2 + 1.0007 \times 10^{-5} * u_x u_{xx} \\
 & - 5.2975 \times 10^{-5} * uu_{xx} - 3.3428 \times 10^{-5} * (u_x)^2 \\
 & + 1.2649 \times 10^{-6} * u_x - 1.5698 \times 10^{-5} * (u)^2 - 1.6640 \times 10^{-6} * u
 \end{aligned}$$

which can be removed by some predefined cut-off value for the coefficient size.

**Table 1**

The variance, feature importance and feature ranking of our dataset for the viscous Burger's equation.

Feature	$u$	$u_x$	$u_{xx}$	$u^2$	$uu_x$	$uu_{xx}$	$u_x^2$	$u_x u_{xx}$	$u_{xx}^2$
Variance	0.21	23	23000	0.06	3.1	5700	11000	$9.0 \times 10^6$	$1.7 \times 10^{10}$
R-Lasso	0.09	0	1	1	1	1	1	1	0
RFE	3	5	2	4	1	6	7	8	9

**Fig. 5.** The logarithm of the cost function for different choices of derivative and non-linear orders  $m$  and  $k$  for the viscous Burger's equation. The true configuration is  $m/k = 2/1$ .

**Remark 4.1.** As the polynomial PDE model is linear we can, of course, use the traditional least squares method with Lasso [40] instead of adding L1 regularization to the optimization problem. In that case we obtain the even sparser model

$$\begin{aligned}
 u_t = & -9.9216 \times 10^{-1} * uu_x + 1.0082 \times 10^{-2} * u_{xx} \\
 & - 2.3627 \times 10^{-9} * (u_{xx})^2 + 1.0129 \times 10^{-5} * u_x u_{xx} \\
 & - 5.4086e \times 10^{-5} * uu_{xx} - 3.2458 \times 10^{-5} * (u_x)^2.
 \end{aligned}$$

The traditional least squares model does not, however, generalize to differential operators of arbitrary complexity or very large datasets.

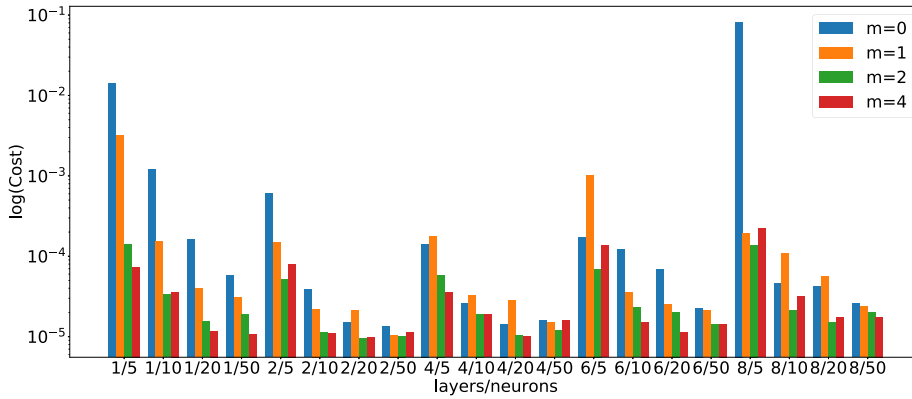
It is common to analyze the input data in order to remove redundant or correlated features. In this case, it is only the terms  $u$ ,  $u_x$ , and  $u_{xx}$  which are independent. A common method is to compute the variance of the input data and remove features with low variance since they are deemed as unimportant. This method does not apply in a PDE context since high order derivatives have lower regularity and hence usually a higher variance, which is clearly shown in Table 1. More sophisticated methods for feature selection include stability analysis via randomized Lasso (R-Lasso) [22], recursive feature elimination (RFE) [12], and Boruta [23]. We include comparisons with the two former methods in Table 1 where we have used the implementations from `scikit-learn` with default parameters. The Boruta method works on ensemble models, such as random forests, and is not suitable in this context. We did, however, try the Boruta method on our dataset with a random forest regressor and we did not obtain any good results. The Boruta method deemed all features as equally important.

We can see from Table 1 that the variance of the features are the opposite of what is expected as the variance grows with the order of the derivative independent of the importance of the feature. By combining R-Lasso and RFE we can get a decent understanding of which features that are important in the dataset.

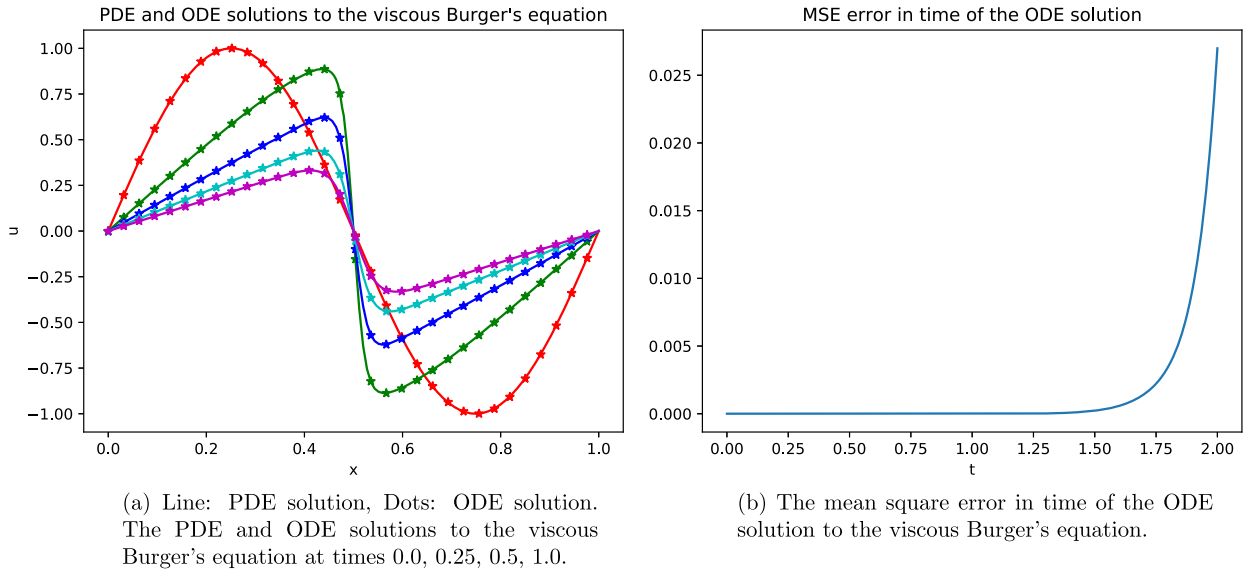
## 5. Model selection

As the polynomial PDE model for the viscous Burger's equation is linear and the optimization problem (2.4) is convex, minimization using standard least squares or gradient based optimization is efficient and model selection can be performed by an exhaustive parameter search. By computing the value of the cost function for different choices of the derivative order  $m$  and non-linear order  $k$ , it is clearly seen when a suitable model has been found. In Fig. 5 we show the logarithm of the cost function for different choices of  $m$  and  $k$ . We can see that the cost function is instantly reduced by several orders of magnitude when a sufficient model has been found.

We can perform a similar study when the PDE is represented by a neural network with different number of layers and neurons. In Fig. 6 we show the value of the cost function for different network designs with different partial derivative orders as input.



**Fig. 6.** The logarithm of the cost function for different network designs and partial derivative orders ( $m$ ) for the viscous Burger's equation.



**Fig. 7.** Comparison between the ODE and PDE solutions of the viscous Burger's equation. The ODE solution is accurate for  $0 \leq t \leq 1$  where we have trained the operator. The ODE operator is, however, unable to extrapolate for  $t \gg 1$ .

The case with 2 hidden layers with 50 neurons in each layer is particularly interesting. In this case we have a low cost even without any partial derivatives as input. Thus for the case  $m = 0$ , the viscous Burger's equation is effectively transformed into an ordinary differential equation (ODE) of the form

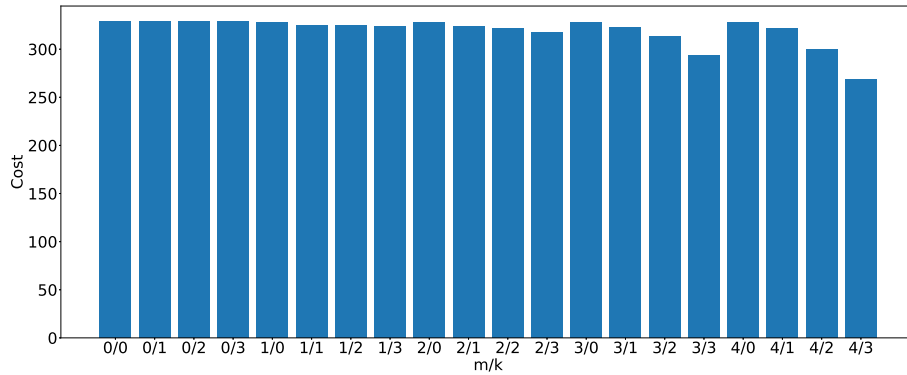
$$\hat{u}_t = \hat{L}(\hat{u}). \quad (5.1)$$

The ODE (5.1) can easily be solved using any time integration method. In Fig. 7 we used standard Runge-Kutta 4(5) from `SciPy` with default settings to integrate the ODE. We can see that the ODE operator gives accurate results for  $0 \leq t \leq 1$  where we have trained the operator. We can also see, unfortunately, that the ODE operator is unable to extrapolate far beyond  $t = 1$  where we have no training data. It is, however, quite remarkable that the dynamics of a second order non-linear PDE can be well approximated by an ODE in the range of the training data and slightly beyond.

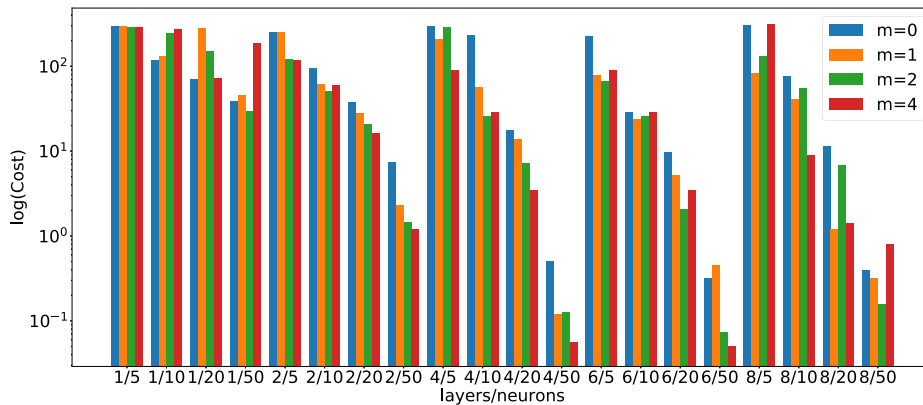
Note that by comparing the result in Fig. 7 with Fig. 2, we can see that the ordinary differential operator is able to better predict the solution further ahead in time than the interpolated solution. The discovered ODE can hence be seen as an efficient extrapolation method for short term predictions.

This method can in the same way be used for model invalidation. Since a PDE model for the temperature distribution is unknown we can perform an exhaustive parameter search to see if a sufficient model can be found. In Fig. 8 we show the value of the cost function for different values of  $m$  and  $k$ , and we can clearly see that there is no sufficient model in this parameter range.

Since no polynomial PDE models for the temperature distribution were found, we can perform the same exhaustive parameter search where we instead vary the number of layers and neurons in each layer when  $\hat{L}$  is represented by a neural network. The results can be seen in Fig. 9 where we represented  $\hat{L}$  by neural networks with 1, 2, 4, 6, 8 hidden layers with



**Fig. 8.** The value of the cost function for different choices of derivative and non-linear orders  $m$  and  $k$  for temperature models. No sufficient polynomial models were found.



**Fig. 9.** The logarithm of the cost function for different network architectures and partial derivative orders ( $m$ ) for temperature models. Some sufficient network models were found.

5, 10, 20, 50 neurons in each layer, respectively, and partial derivatives of order  $m = 0, 1, 2, 4$  as input. We can see that the cost drops several orders of magnitude for certain configurations which indicate that sufficient models have been found. We can also see that even in this complicated case, there are some ODE models which appears to capture the dynamics.

Similarly to viscous Burger's case, we use the ODE operator with 6 layers and 50 neurons in each layer to compute the mean square error in time for the ODE solution using Runge-Kutta 4(5). In this case, the ODE operator is trained on data from the first week in July 2016 ( $0 \leq t \leq 1$ ) and evaluated on both the first and second week ( $0 \leq t \leq 2$ ) to test the prediction performance. As in the viscous Burger's case, we can see in Fig. 10 that the ODE operator is fairly accurate in the region where training data is available but is unable to extrapolate far beyond the training data. However, the operator is able to remain accurate up to time  $t = 1.25$  which amounts to quarter of a week in physical time.

The simulation shown in Fig. 10 of the Swedish temperature distribution over a two week period using the ODE operator takes only a fraction of a second on a laptop. We hence believe that by incorporating more quantities in the measurements, it is possible to discover a system of ODEs which can be used to obtain both fast and accurate short-time predictions.

## 6. Summary and conclusions

We have used deep artificial neural networks to discover partial differential equations from data sets consisting of measurements of physical quantities of interest. The quantities of interest are both artificial from known model PDEs, as well as true measurement data from weather stations.

In general, the physical domain is non-trivial and data transformations are necessary to bring the problem into a range where machine learning algorithms perform well. These data transformations amounts to coordinate transformations in the discovered PDEs and it is hence important that all data transformations are recorded such that the discovered PDEs can be transformed back into physical space. We have shown examples of general data transformations and the common shift and scale transformation in particular.

The discovered PDE operator is not unique for any given data set. We performed parameter searches to discover a range of operators that describes a PDE which is generating our data set. We found that the dynamics of the non-linear, second order viscous Burger's equation could also be well approximated by an ODE which was automatically discovered. We also found an ODE for a 2D temperature distribution model which shows interesting properties for further research. The ODE

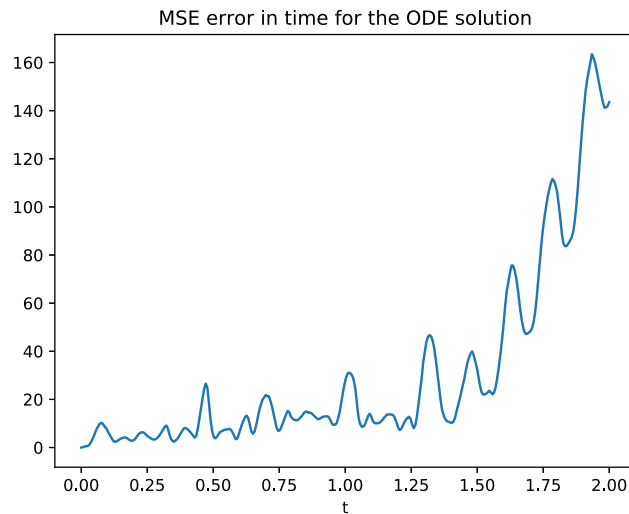


Fig. 10. The mean square error in time of the ODE temperature model.

operators we found are accurate in the region of the training data and are able to extrapolate slightly beyond the training data. The benefit of the ODE models is that they can be solved in fractions of a second on a laptop, compared to the PDE models which require substantial computational resources. We believe that automatically derived ODEs from observed data can be used for efficient extrapolation for short-term predictions. We saw that the ODEs were as accurate as the neural network interpolation of the measured data within the range of the data, but they were able to extrapolate much further than the interpolated solution.

### Acknowledgements

Some of the computations were performed on resources provided by The Swedish National Infrastructure for Computing (SNIC) through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Project SNIC 2017/7-131.

The authors were partially supported by a grant from the Göran Gustafsson Foundation for Research in Natural Sciences and Medicine.

### References

- [1] Comsol multiphysics, <http://www.comsol.com>. (Accessed 22 May 2018).
- [2] 10 key marketing trends for 2017 and ideas for exceeding customer expectations, 2017, IBM, online. (Accessed 23 April 2018).
- [3] M. Abadi, et al., TensorFlow: large-scale machine learning on heterogeneous systems, software available from [tensorflow.org](https://www.tensorflow.org), 2015.
- [4] J. Berg, K. Nyström, Neural network augmented inverse problems for PDEs, arXiv e-prints, Dec. 2017.
- [5] S.L. Brunton, J.L. Proctor, J.N. Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems, *Proceedings of the National Academy of Sciences* (2016).
- [6] G. Cybenko, Approximation by superpositions of a sigmoidal function, *Mathematics of Control, Signals, and Systems* 2 (4) (Dec. 1989) 303–314.
- [7] D.C.D.R. Fernández, J.E. Hicken, D.W. Zingg, Review of summation-by-parts operators with simultaneous approximation terms for the numerical solution of partial differential equations, *Computers & Fluids* 95 (2014) 171–196.
- [8] R. Fletcher, *Practical Methods of Optimization*, 2nd edition, Wiley, 1987.
- [9] H. Garnier, L. Wang, *Identification of Continuous-Time Models from Sampled Data*, Springer, London, 2008.
- [10] L. Guo, S. Billings, Identification of partial differential equation models for continuous spatio-temporal dynamical systems, *IEEE Transactions on Circuits and Systems II: Express Briefs* 53 (8) (Aug. 2006) 657–661.
- [11] L. Guo, S. Billings, D. Coca, Identification of partial differential equation models for a class of multiscale spatio-temporal dynamical systems, *International Journal of Control* 83 (1) (Aug. 2009) 40–48.
- [12] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, *Machine Learning* 46 (1/3) (2002) 389–422.
- [13] J. Hadamard, Sur les problèmes aux dérivées partielles et leur signification physique, *Princeton University Bulletin* 13 (1902) 49–52.
- [14] K. Hornik, Approximation capabilities of multilayer feedforward networks, *Neural Networks* 4 (2) (1991) 251–257.
- [15] K. Hornik, M. Stinchcombe, H. White, Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks, *Neural Networks* 3 (5) (1990) 551–560.
- [16] E. Jones, T. Oliphant, P. Peterson, et al., SciPy: open source scientific tools for Python, <http://www.scipy.org>, 2001. (Accessed 12 December 2017).
- [17] A. Juditsky, H. Hjalmarsson, A. Benveniste, B. Delyon, L. Ljung, J. Sjöberg, Q. Zhang, Nonlinear black-box models in system identification: mathematical foundations, *Automatica* 31 (12) (Dec. 1995) 1725–1750.
- [18] Y. LeCun, L. Bottou, G.B. Orr, K.R. Müller, Efficient backprop, in: *Neural Networks: Tricks of the Trade*, Springer, 1998, pp. 9–50.
- [19] D.C. Liu, J. Nocedal, On the limited memory BFGS method for large scale optimization, *Math. Program.* 45 (1) (Aug. 1989) 503–528.
- [20] L. Ljung, T. Glad, *Modeling of Dynamic Systems*, Prentice Hall, 1994.
- [21] A. Logg, K.-A. Mardal, G.N. Wells, et al., *Automated Solution of Differential Equations by the Finite Element Method*, Springer, 2012.
- [22] N. Meinshausen, P. Bühlmann, Stability selection, *J. R. Stat. Soc., Ser. B, Stat. Methodol.* 72 (4) (July 2010) 417–473.

- [23] R. Nilsson, J.M. Peña, J. Björkegren, J. Tegnér, Consistent feature selection for pattern recognition in polynomial time, *J. Mach. Learn. Res.* 8 (Dec. 2007) 589–612.
- [24] J.P. Crutchfield, B.S. McNamara, Equations of motions a data series, *Complex Syst.* 1 (Jan. 1987) 417–452.
- [25] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in PyTorch, in: *NIPS-W*, 2017.
- [26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in Python, *J. Mach. Learn. Res.* 12 (2011) 2825–2830.
- [27] M. Raissi, Deep hidden physics models: deep learning of nonlinear partial differential equations, *arXiv e-prints*, Jan. 2018.
- [28] M. Raissi, G.E. Karniadakis, Hidden physics models: machine learning of nonlinear partial differential equations, *J. Comput. Phys.* 357 (2018) 125–141.
- [29] M. Raissi, P. Perdikaris, G.E. Karniadakis, Machine learning of linear differential equations using Gaussian processes, *J. Comput. Phys.* 348 (2017) 683–693.
- [30] M. Raissi, P. Perdikaris, G.E. Karniadakis, Physics informed deep learning (part I): data-driven solutions of nonlinear partial differential equations, *arXiv e-prints*, Nov. 2017.
- [31] M. Raissi, P. Perdikaris, G.E. Karniadakis, Physics informed deep learning (part II): data-driven discovery of nonlinear partial differential equations, *arXiv e-prints*, Nov. 2017.
- [32] M. Raissi, P. Perdikaris, G.E. Karniadakis, Numerical Gaussian processes for time-dependent and nonlinear partial differential equations, *SIAM J. Sci. Comput.* 40 (1) (2018) A172–A198.
- [33] S.H. Rudy, S.L. Brunton, J.L. Proctor, J.N. Kutz, Data-driven discovery of partial differential equations, *Sci. Adv.* 3 (4) (2017).
- [34] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, *Nature* 323 (6088) (Oct. 1986) 533–536.
- [35] H. Schaeffer, Learning partial differential equations via data discovery and sparse optimization, *Proc. R. Soc. A, Math. Phys. Eng. Sci.* 473 (2197) (Jan. 2017) 20160446.
- [36] H. Schaeffer, R. Caflisch, C.D. Hauck, S. Osher, Sparse dynamics for partial differential equations, *Proc. Natl. Acad. Sci.* 110 (17) (Mar. 2013) 6634–6639.
- [37] J. Sjöberg, Q. Zhang, L. Ljung, A. Benveniste, B. Delyon, P.-Y. Glorennec, H. Hjalmarsson, A. Juditsky, Nonlinear black-box modeling in system identification: a unified overview, *Automatica* 31 (12) (Dec. 1995) 1691–1724.
- [38] T. Söderström, P. Stoica, *Modeling of Dynamic Systems*, Prentice Hall, 1989.
- [39] M. Svärd, J. Nordström, Review of summation-by-parts schemes for initial-boundary-value problems, *J. Comput. Phys.* 268 (2014) 17–38.
- [40] R. Tibshirani, Regression shrinkage and selection via the Lasso, *J. R. Stat. Soc., Ser. B, Methodol.* 58 (1) (1996) 267–288.
- [41] P. Young, Parameter estimation for continuous-time models—a survey, *Automatica* 17 (1) (1981) 23–39.