Report for STM 2015:

Transferring Information from Data to Model

Fabio Vallone[1,2]

1. Institute of Biophysics, CNR-National Research Council, 56124 Pisa, Italy

2. Translational Neural Engineering Area, The Biorobotics Institute,
Scuola Superiore Sant'Anna, 56026 Pisa, Italy

(Dated: January 20, 2016)

I. INTRODUCTION

The present report is based on my research activity at the Department of Physics of the University of California San Diego (UCSD), La Jolla, California, USA. I have been hosted in the group of Prof. H. Abarbanel, which is one of the pioneer of the statistical data assimilation methods using path integral techniques. During this research period I learned the main aspects of the statistical data assimilation problem and I acquired the formal and numerical recipes to solve the problem. In the following, I presented the methods that I learned with relevant application to neural dynamics.

Making and testing models of observed time series from dynamical systems is the forefront of nonlinear dynamics research. In this report, we studied theoretical problems in nonlinear time series analysis and modeling [1–3]. We focused on modern methods of transferring information from data to models, this is *data assimilation* (see [1] and references therein) .

The problem of data assimilation is quite simple to state: during a measurement window $[0, T]$, one makes observations at discrete time steps $[t_0 = 0, t_1, ..., t_m = T]$ of some physical quantities of a dynamical system. Given a mathematical model describing the observed dynamical system, we want to determine the parameters entering into the model and to estimate both the observed and *unobserved* physical quantities of the system during the measurement time interval $[0, T]$.

The model typically has a lot of state variables, not all of which we are able to observe, and it has some physical parameters we may to know. Furthermore, if we are able to get good estimations of parameters and full state variables (observed and unobserved), we can use the model to *predict* the behavior of the sytem for $t > T$, thus the prediction can be a test of the validity of the model [1].

Adopting predictions as a metric to test model is not a simple task because a variety of reasons can produce a wrong prediction. For example, due to a lot of amount of noise in the data the behavior of the physical systems cannot be recovered by the model. Moreover, an insufficient amount of data can result to a flawed data assimilation procedure or the model is incorrect.

The problem of noisy data together with model errors makes the data assimilation a statistical problem. Indeed, we have to calculate a probability distribution for the state of the model, conditioned on the observations. One must start with a, known or unknown, initial distribution of states at $t = 0$ and then propagate this probability distribution using the dynamical rule of our model to the first measurement time t_1. At t_1 we require a rule letting us know how information in the measurement at that time influences the state distribution function at that time. Then, using the dynamical model, we need to propagate that distribution function to the next measurement time $t_2 > t_1$ and apply our information transfer rule and so forth until we reach T (the end of the observation window). These steps characterize the statistical data assimilation techniques and are done using a path integral formalism allowing us to give an integral representation of the solution to the data assimilation problem.

In the following sections we formally stated the problem and we gave an overview of the method used to solve the statistical data assimilation problem and its applications. In particular, we discussed a physical relevant example on neural dynamics, i.e. an Hodgkin-Huxley model of a neuron with potassium and sodium ionic channels [1, 4, 5] , that is called the NaKL model. Finally, we gave an outlook on a model of the telencephalic nucleus HVC that contains neurons playing a crucial role in birdsong system[6–8] .

II. METHODS

A. Dynamical systems and data assimilation

A dynamical model of a physical system can be written in terms of a set of differential equations in the following form:

$$\frac{dx_a(t)}{dt} = F_a\left(\mathbf{x}(t), \mathbf{p}\right), \tag{1}$$

where $a = 1, ..., D$ with D defining the dimension of system and $\mathbf{x}(t) = [x_1(t), x_2(t), ..., x_D(t)]$ is a vector describing the state of the system at time t. $\mathbf{F}\left(\mathbf{x}(t)\right)$ is a dynamical vector field governing the dynamics of our system and $\mathbf{p}$ are the N_p parameters entering into the model. The corresponding discrete form of the dynamical Eq. 1 is

$$x_a(t_{n+1}) = f_a\left(\mathbf{x}(t_n), \mathbf{p}\right). \tag{2}$$

The information we wish to transfer to this model resides in the L measurements $\mathbf{y}(t_n) = [y_1(t_n), y_2(t_n), ..., y_L(t_n)]$ made at each time t_n within an observation window

$[t_0 = 0, t_1, ..., t_m = T]$. To connect the measurements $y_l(t_n)$ with the solution of the dynamical Eq. 2, we must specify a measurement function h, as follows $h_l(\mathbf{x}(t_n)) = y_l(t_n)$.

The use of measurements in the observation time window $[0, T]$ to estimate the parameters $\mathbf{p}$ and *unknown* states *completes* the model and allow us to *test* or *validate* the model using the predictions for $t > T$ where a selected metric compares new observations $\mathbf{y}(t > T)$ to new model outpus $h_l(\mathbf{x}(t_n))$.

Typically, the measurement are sparse, i.e. $L << D$, and we must estimate both the full (observed an unobserved) D-dimensional state $\mathbf{x}$ and N_p parameters of the model. This limitation makes the data assimilation problem very difficult [1, 9] .

It is worth remarking that, in many cases it is found that the trajectory of the measured variable coincide quite well with the data within the observation window. However, the quality of the estimation of parameters and unobserved states cannot be ascertained without further tests. In this way, the *true test* of the assimilation procedure is comparison of the *predictions* of the state variables for $t > T$. Indeed, it is often found that excellent estimations lead to unsatisfactory predictions of the measured states [1].

B. Path Integral representation for statistical data assimilation problem

As discussed in the Introduction, experimental measures are always noisy and the model describing the observed physical system has always errors, thus the data assimilation problem becomes a statistical problem. We do not enter into the detail of the general formulation of the statistical data assimilation problem (see [1] and references therein for details), however we write the general solution to the problem and we comment it in a intuitive way. For brevity reasons, we define $\mathbf{y}(t_n) = \mathbf{y}(n)$ and $\mathbf{x}(t_n) = \mathbf{x}(n)$.

The main ingredient is the evaluation of the conditional probability $P(\mathbf{X}|\mathbf{Y})$, where $\mathbf{Y} = [\mathbf{y}(m), \mathbf{y}(m-1), ..., \mathbf{y}(0)]$ are the observations and $\mathbf{X} = [\mathbf{x}(m), \mathbf{x}(m-1), ..., \mathbf{x}(0)]$ is the collection of states of the system during the observation window.

It can be shown that $P(\mathbf{X}|\mathbf{Y}) = \exp[-A_0(\mathbf{X}, \mathbf{Y})]$, where $A_0(\mathbf{X}, \mathbf{Y})$ is called the action and is equal to

$$A_0(\mathbf{X}, \mathbf{Y}) = -\sum_{n=0}^{m} CMI(\mathbf{x}(n), \mathbf{y}(n)|\mathbf{Y}(n-1)) - \sum_{n=0}^{m-1} \ln\left[P(\mathbf{x}(n+1)|\mathbf{x}(n))\right] - \ln\left[P(\mathbf{x}(0))\right]. \quad (3)$$

Since usually we do not have prior information of the initial distribution of the states, $P(\mathbf{x}(0))$ is assumed to be a uniform distribution and the last term of the above equation can be ignored as an additive constant. The term $\ln\left[P(\mathbf{x}(n+1)|\mathbf{x}(n))\right]$ contains the information on the transition from a state at time t_n to a state at time t_{n+1}. On the other hand, the first term of the above equation is associated with the measurement during the observation window. It can be expressed as a conditional mutual information in the following way

$$CMI(\mathbf{x}(n),\mathbf{y}(n)|\mathbf{Y}(n-1)) = \ln\left[\frac{P(\mathbf{x}(n),\mathbf{y}(n)|\mathbf{Y}(n-1))}{P(\mathbf{x}(n)|\mathbf{Y}(n-1))P(\mathbf{y}(n)|\mathbf{Y}(n-1))}\right] \tag{4}$$

where $\mathbf{Y}(n{-}1) = [\mathbf{y}(n), \mathbf{y}(n-1), ..., \mathbf{y}(0)]$ is the collection of the measurements up to time t_n. This term contains the additional information transferred from the current measurement $\mathbf{y}(n)$ to the model $\mathbf{x}(n)$, conditioned on the past measurements in $\mathbf{Y}(n-1)$.

Our main interest is in the calculation of conditional expectation values of statistical quantities expressed as a function of $\mathbf{X}$, i.e. $G(\mathbf{X})$, that can be expressed as a path integral of the following form:

$$E\left[G(\mathbf{X})|\mathbf{Y}\right] = \frac{\int d\mathbf{X}\, G(\mathbf{X})\exp[-A_0(\mathbf{X},\mathbf{Y})]}{\int d\mathbf{X}\,\exp[-A_0(\mathbf{X},\mathbf{Y})]}. \tag{5}$$

This is an high dimensional $D(m+1)$ integral along the path of the model state through the observation window $[t_0 = 0, t_1, ..., t_m = T]$. The expected path through the space of paths during the observation window comes from selecting $G(\mathbf{X}) = \mathbf{X}$. The computation of the RMS variation about this expected path is performed with the information contained in $G(\mathbf{X}) = [\mathbf{x}(0)^2, \mathbf{x}(1)^2, ..., \mathbf{x}(m)^2]$. Other moments about the expected path are evaluated in the same way using different powers of the $\mathbf{x}(n)$.

C. Calculating the Path Integral: Annealing procedure

In the spirit of Lagrangian dynamics and perturbative field theory, one may more systematically approach the problem of calculating the integral in Eq. 5 by expanding around stationary paths, that is, utilizing Laplace's method (see [1] and references therein) . The computational difficulty of the problem is thus shifted to one of nonlinear optimization - to finding the minima of $A_0(\mathbf{X},\mathbf{Y})$.

If one assumes both measurement noises and model error are independent and gaussian, the action A_0 in Eq. 3 has the format of

$$A_0(\mathbf{X},\mathbf{Y}) = \sum_{n=0}^{m} \frac{R_m(n)}{2} \sum_{l=1}^{L} [x_l(n) - y_l(n)]^2 + \frac{R_f}{2} \sum_{n=0}^{m-1} \sum_{a=1}^{D} [x_a(n+1) - f_a(\mathbf{x}(n))]^2. \quad (6)$$

where R_m and R_f are the inverse of variances for the measurement noises and model error, respectively.

The annealing method (see [5] for details) is based on the observation that the minima solution X^q of A_0 at $R_f = 0$ is $x_l(n) = y_l(n)$, the other $D - L$ components of the model state vector are undetermined, and the solution is degenerate. As we increase R_f, the action levels split, and depending on R_m, R_f, L and the precise form of the dynamical vector field $\mathbf{f}(\mathbf{x})$, there will be 1,2,... minima of A_0.

After identifying the global minima and other local minima of A_0, we can employ laplace method to approximate the expected value $E\left[G(\mathbf{X})|\mathbf{Y}\right]$ of a function $G(\mathbf{X})$, i.e

$$E\left[G(\mathbf{X})|\mathbf{Y}\right] = \frac{\int d\mathbf{X}\, G(\mathbf{X}) \exp[-A_0(\mathbf{X},\mathbf{Y})]}{\int d\mathbf{X}\, \exp[-A_0(\mathbf{X},\mathbf{Y})]} \approx G(\mathbf{X^0}). \quad (7)$$

plus exponentially small corrections.

III. APPLICATION OF DATA ASSIMILATION ON NEURONAL DYNAMICS

In this section we give an example of the application of data assimilation technique to an Hodgkin-Huxley model of a neuron with potassium an sodium channel: the NaKL model.

We started to test our data assimilation method via *twin experiments*. In twin experiment the data are generated by a known model (in our case the NaKL model) and the statistical data assimilation methods are tested on this model. Twin experiments are also very important as they allow us to address questions such as how many measurements are required for the accuracy of state and parameter estimations [1, 9] . They can be used to identify which measurements to make. They permit us to ask how frequently in time this number of measurements should be performed in order to make accurate estimations of model parameters and model states. Twin experiments are also useful as a design tool for experiments as they may indicate properties of the stimulus or forcing of the experimental system in order to explore the full dynamical range of its response [10].

A. NaKL model of neuron

The NaKL model based on Hodgkin-Huxley equations describes the membrane potential of a neuron consisted by potassium an sodium ion channels. The NaKL model has four state variables and 25 fixed parameters. The dynamical equation for the membrane potential $V(t)$ is

$$\frac{dV(t)}{dt} = g_{\mathrm{L}}(E_{\mathrm{L}} - V(t)) + g_{\mathrm{K}} n(t)^4 (E_{\mathrm{K}} - V(t)) + g_{\mathrm{Na}} m(t)^3 h(t)(E_{\mathrm{Na}} - V(t)) + I_{\mathrm{injected}} \tag{8}$$

where $g_{\mathrm{K}}, E_{\mathrm{K}}$ and $g_{\mathrm{Na}}, E_{\mathrm{Na}}$ are the conductances and the reversal potential for the potassium and sodium channels, respectively. g_L, E_L are the conductances and the reversal potential for the leak channel. I_{injected} is an injected external current.

The gating variables $a = m, h, n$ describing the activation and inactivation of a channel are written in the following form

$$\frac{da(t)}{dt} = \frac{a_\infty(V(t)) - a(t)}{\tau_a(V(t))}, \qquad a = m, h, n \tag{9}$$

$$a_\infty(V(t)) = \frac{1}{2}\left(1 + \tanh\left(\frac{V(t) - \theta_a}{\sigma_a}\right)\right), \tag{10}$$

$$\tau_a(V(t)) = \tau_{a0} + \tau_{a1}\left(1 - \tanh^2\left(\frac{V(t) - \theta_{\tau a}}{\sigma_{\tau a}}\right)\right). \tag{11}$$

These dynamical equations are used to produce virtual data $\mathbf{x}_{\mathrm{data}}(t) = [V_{\mathrm{data}}(t), m_{\mathrm{data}}, h_{\mathrm{data}}, n_{\mathrm{data}}]$ that were analyzed in a twin experiment.

B. Twin experiments with NaKL model neuron

According to nowadays experimental technology, in our twin experiment the membrane potential $V_{\mathrm{data}}(t)$ was the observed variables, whereas the gating variables $m_{\mathrm{data}}, h_{\mathrm{data}}, n_{\mathrm{data}}$

were the *unobserved* state variables.

The dynamical Eq.s 8,9 of the NaKL model are integrated using a Runge-Kutta fourth order method [11], with a time step integration of $\Delta t = 0.02$ms, corresponding to frequency sampling of $f_s = 1/\Delta t = 50$kHz. This time step was chosen in order to achieve a good data assimilation procedure, indeed it was found that sampling rate lower than 50 kHz did not allow a good estimation of the parameters and variables of the model [1]. The external injected current in Eq. 8 is shown in the top panel of Fig. 1. This is a chaotic current that is able to stimulate all the dynamical range of the system. It was shown that the usual step currents are not sufficient to provide good estimation of the parameters and state of the system as well as predictions [1]. Finally, gaussian noise was added to the data of the membrane potential, to reproduce an experimental error of ±1 mV. Thus, we produced a virtual experimental data set of the membrane potential $V_{\text{data}}(t)$ (see bottom panel of Fig. 1). Only $V_{\text{data}}(t)$ is presented to the the NaKL model in our twin experiments (according to our previous notation $V_{\text{data}}(t) = \mathbf{y}(t)$) and the path integral was calculated using the annealing procedure (see Sec.II C) on the first 3000 data points. The numerical minimization problem was performed using the interior-point algorithm provided by the open source software IPOPT [1, 5] , utilizing the ma57 linear solver library, on a standard desktop computer. Fig. 2 shows the values of the action A_0 for the paths with minimum action at fixed R_f. The paths corresponding to the minimum asymptotic value of A_0 is chosen as the minimum action path. In our example $R_f = \{R_{f0}\alpha^{\beta}\}$, where $\alpha = 2, \beta = 0, 1, \ldots, 30$, $R_{f0} = 10^{-3}$, the path with minimum action corresponded to value of $\beta = 30$. The results for the parameter estimation are very good as can be seen by the value represented in Table I. The result for the estimated membrane voltage is shown as blue line in Fig.3, whereas the green line indicates the prediction that is in excellent agreement with the generated data (red line in Fig.3). Also for the *unobserved* gating variables of the ion channels we achieved very good estimation and prediction (see Fig.s4,5). Overall, these results indicate that the statistical data assimilation methods is correct showing its powerful application both on the completion and prediction of a nonlinear dynamical system such as the NaKL model. The twin experiment strongly suggested a design for experiment in a single neuron in which an external chaotic current is applied and the frequency sampling of the potential recording must be 50KHz. These requirements are crucial for the success of the statistical data assimilation technique.

IV. CONCLUSION AND OUTLOOK

The experience at the Department of Physics of the UCSD was highly formative allowing me to enter in a new issue such as the statistical data assimilation problem. In this report, we reviewed the main aspect of the statistical data assimilation problem and we gave the formal and numerical recipes to solve the problem. We have shown its applications on a relevant biophysical model of an NaKL neuron described by Hodgkin-Huxley equations. The twin experiment performed using the NakL model was useful as a design for experiment in a single neuron obtaining requirements on the frequency sampling and on the external injected current. Moreover, we obtained nice estimations and prediction of both the observed and *unobserved* state variables allowing us to complete the model and to explore the full dynamics of the system.

However, twin experiment are based on the fact that we know the model generating the data, a situation that is not possible in real data experiment in which the dynamical model generating the system is unknown. Currently, in [12] statistical data assimilation is performed on real data recorded from HVC neurons that play a crucial role in birdsong system [6–8].

After the complete characterization of the biophysical properties of individual neurons, the main goal is to build a model of the whole HVC nucleus in the avian song system where thousands of neurons are interacting through synapses [6–8]. A relevant problem in building such large network resides on the missing information about the connectivity of the network. In fact, due to the sparseness of the data it is not possible to infer the exact geometry of the network. Previous attempt to build such model (see [7]) are based on several hypothesis which can be questionable, and the testing of its validity requires very difficult experimental techniques [6]. In this framework, the statistical data assimilation methods are very promising to furnish the criteria to require which type of measurements we need to infer the connectivity of the network and then to complete the model and to test it with the prediction metrics.

ACKNOWLEDGEMENT

This work was supported by the Short-Term Mobility project of the National Research Council of Italy-CNR. The analysis were done at the Department of Physics of the University of California San Diego (UCSD), La Jolla, California, USA. A special thanks to Henry Abarbanel and all his group for the kind hospitality and the many discussions. Finally, a thanks to Dr. Angelo Di Garbo for reading the manuscript and for his comments to improve the quality of the report.

[1] H. Abarbanel, *Predicting the Future: Completing Models of Observed Complex Systems* (Springer, 2013) p. 238

[2] H. Abarbanel, *Analysis of Observed Chaotic Data* (Springer Science & Business Media, 1996) p. 272

[3] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge University Press, 2004) p. 369

[4] B. A. Toth, M. Kostuk, C. D. Meliza, D. Margoliash, and H. D. I. Abarbanel, Biol. Cybern. **105**, 217 (2011)

[5] J. Ye, D. Rey, N. Kadakia, M. Eldridge, U. I. Morone, P. Rozdeba, H. D. I. Abarbanel, and J. C. Quinn, Phys. Rev. E. Stat. Nonlin. Soft Matter Phys. **92**, 052901 (2015)

[6] G. Kosche, D. Vallentin, and M. A. Long, J. Neurosci. **35**, 1217 (2015)

[7] L. Gibb, T. Q. Gentner, and H. D. I. Abarbanel, J. Neurophysiol. **102**, 1748 (2009)

[8] R. H. R. Hahnloser, A. A. Kozhevnikov, and M. S. Fee, Nature **419**, 65 (2002)

[9] D. Rey, M. Eldridge, U. Morone, H. D. I. Abarbanel, U. Parlitz, and J. Schumann-Bischoff, Phys. Rev. E. Stat. Nonlin. Soft Matter Phys. **90**, 062916 (2014)

[10] N. Kadakia, E. Armstrong, D. Breen, U. Morone, A. Daou, D. Margoliash, and H. D. I. Abarbanel, In preparation (2016)

[11] W. H. Press, *Numerical Recipes 3rd Edition: The Art of Scientific Computing* (Cambridge University Press, 2007) p. 1235

[12] D. Breen, et al. In preparation (2016)

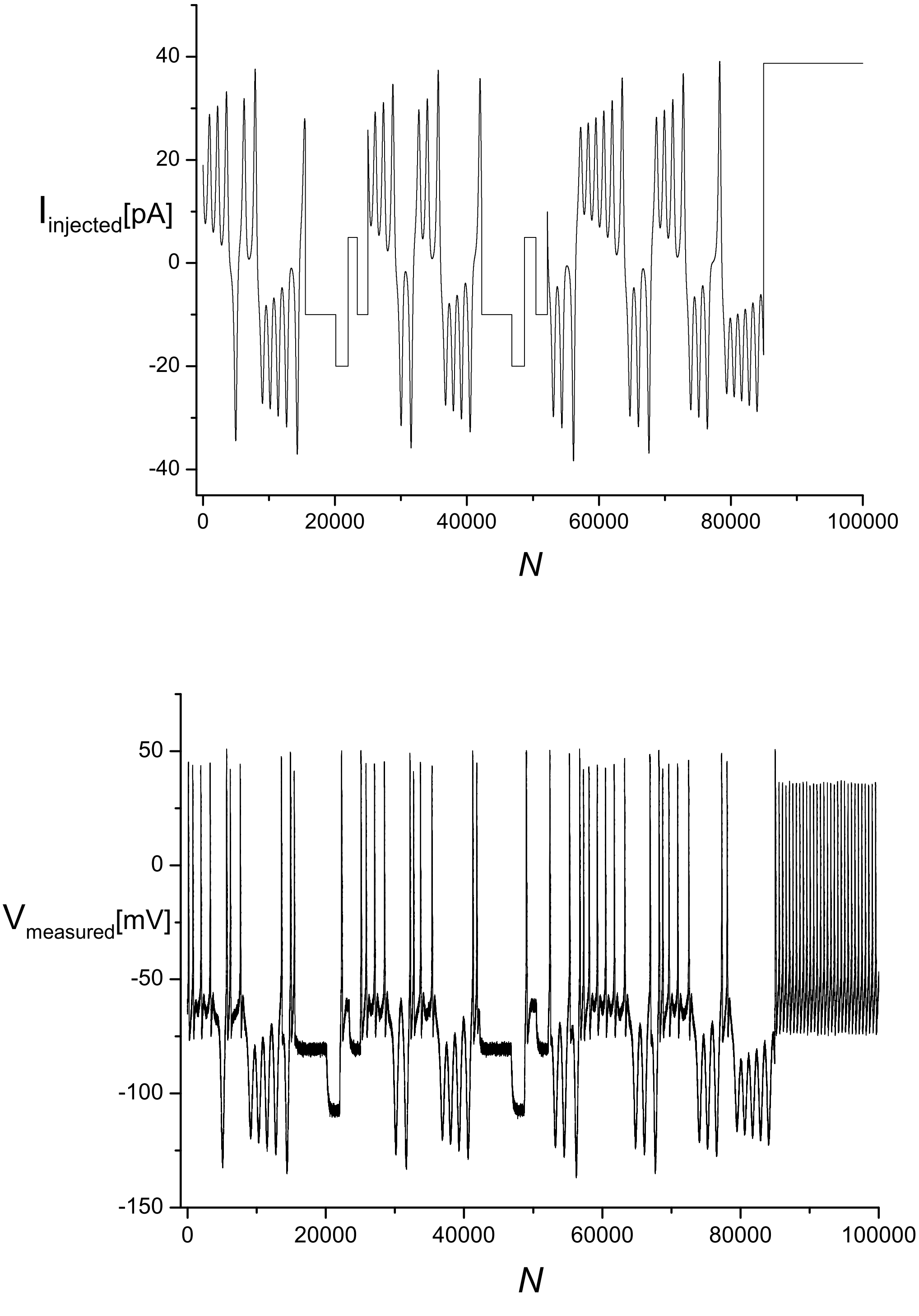

FIG. 1. (Top panel) The chaotic injected current used to stimulate the neuron of the NaKL model. (Bottom panel) Membrane potetial measured in twin experiment.

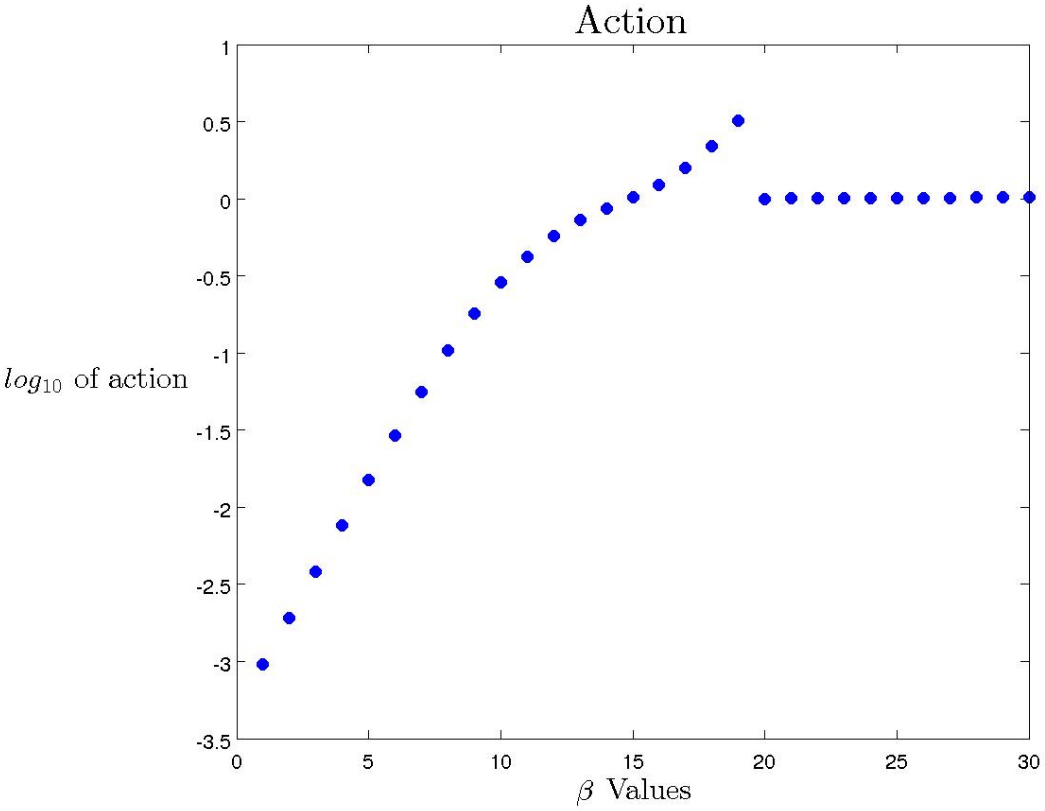

FIG. 2. Action level values over increasing values of $R_f = \{R_{f0}\alpha^{\beta}\}$, where $\alpha = 2, \beta = 0, 1, \ldots, 30$, $R_{f0} = 10^{-3}$. Continue this process until the lowest action level path X^0 produces a $A_0(X^0)$ near expected value, which can be identified from our knowledge of measurement noises. With our choice of parameters the limiting value of the action is 1.

Parameter name	**Parameter value in the data**	**Estimated value of the parameter**
g_{Na}	120 mS/cm^2	108.3718 mS/cm^2
E_{Na}	50 mV	49.98036 mV
g_{K}	20 mS/cm^2	21.10989 mS/cm^2
E_{K}	-77 mV	-77.09009 mV
g_{L}	0.30 mS/cm^2	0.302808 mS/cm^2
E_{L}	54 mV	-54.04558 mV
$\theta_{\mathrm{m}} = \theta_{\tau\mathrm{m}}$	-40 mV	-40.23915 mV
$\sigma_{\mathrm{m}} = \sigma_{\tau\mathrm{m}}$	15 mV	14.93786 mV
τ_{m0}	0.1 ms	0.094947 ms
τ_{m1}	0.4 ms	0.411992 ms
$\theta_{\mathrm{h}} = \theta_{\tau\mathrm{h}}$	-60 mV	-59.43191 mV
$\sigma_{\mathrm{h}} = \sigma_{\tau\mathrm{h}}$	15 mV	14.24096 mV
τ_{h0}	1 ms	1.032077 ms
τ_{h1}	7 ms	7.760693 ms
$\theta_{\mathrm{n}} = \theta_{\tau\mathrm{n}}$	-55 mV	-54.5152 mV
$\sigma_{\mathrm{n}} = \sigma_{\tau\mathrm{h}}$	30 mV	30.49059 mV
τ_{n0}	1 ms	1.059519 ms
τ_{n1}	5 ms	4.965534 ms

TABLE I. Parameter estimation for the NaKL model. The estimation are really close to the value of the parameters entering into the model generating the data.

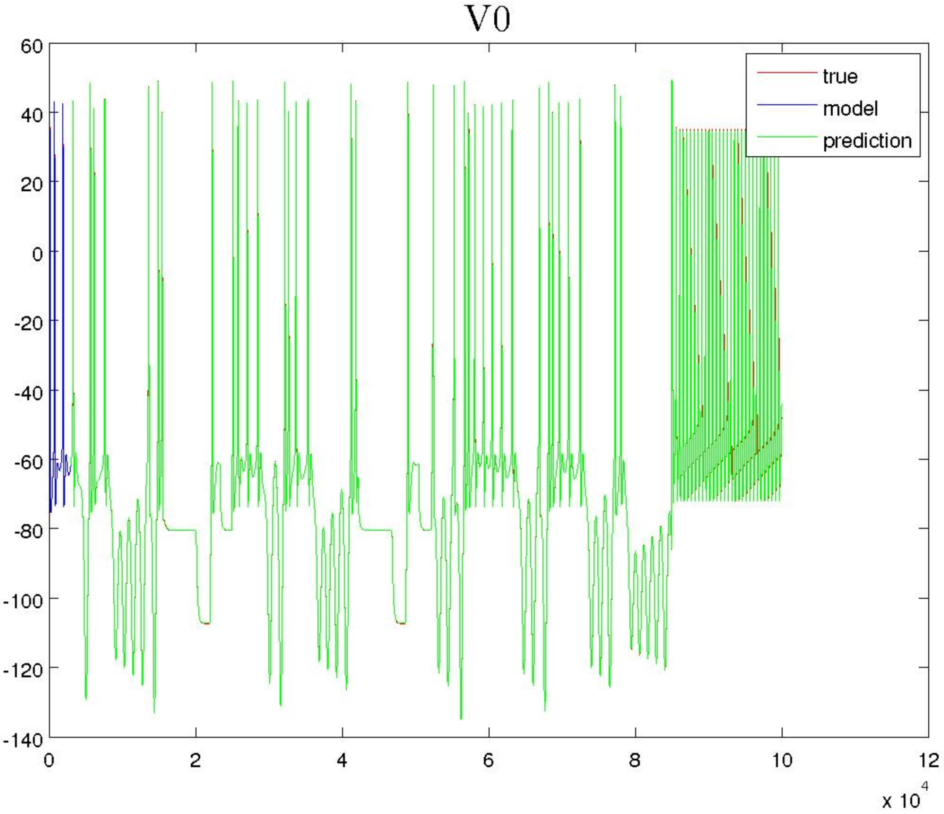

FIG. 3. Results for the *observed* membrane potential. The blue line represents the estimated membrane potential during the assimilation window (3000 data points), whereas the prediction and the true data are represented by the green and red line, respectively. As can be seen by the figure, the prediction are excellent indicating the validity of the data assimilation methods.

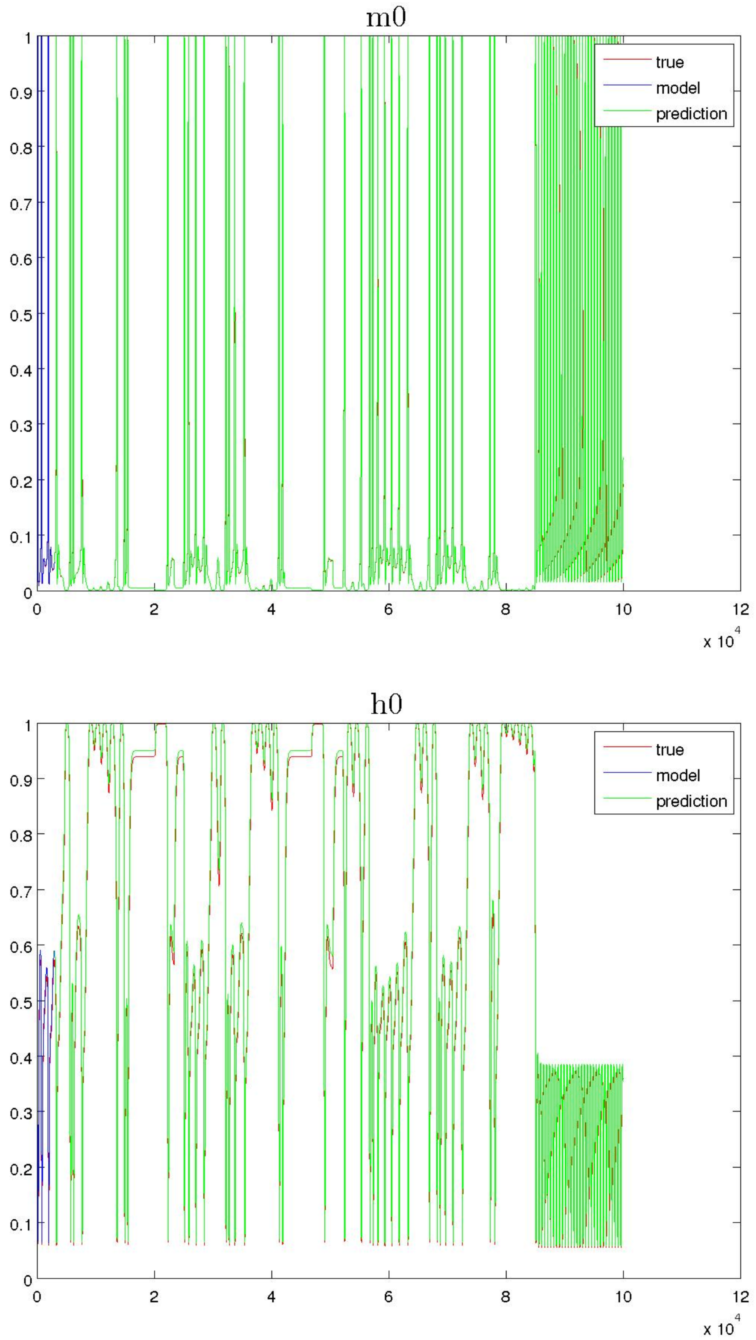

FIG. 4. Results for the *unobserved* gating variables of the sodium channel. The blue lines represent the estimated gating variables m (top panel) and h (bottom panel) during the assimilation window (3000 data points), whereas the prediction and the true data are represented by the green and red line, respectively. Also in this case the predictions are excellent indicating the validity of the data assimilation methods.

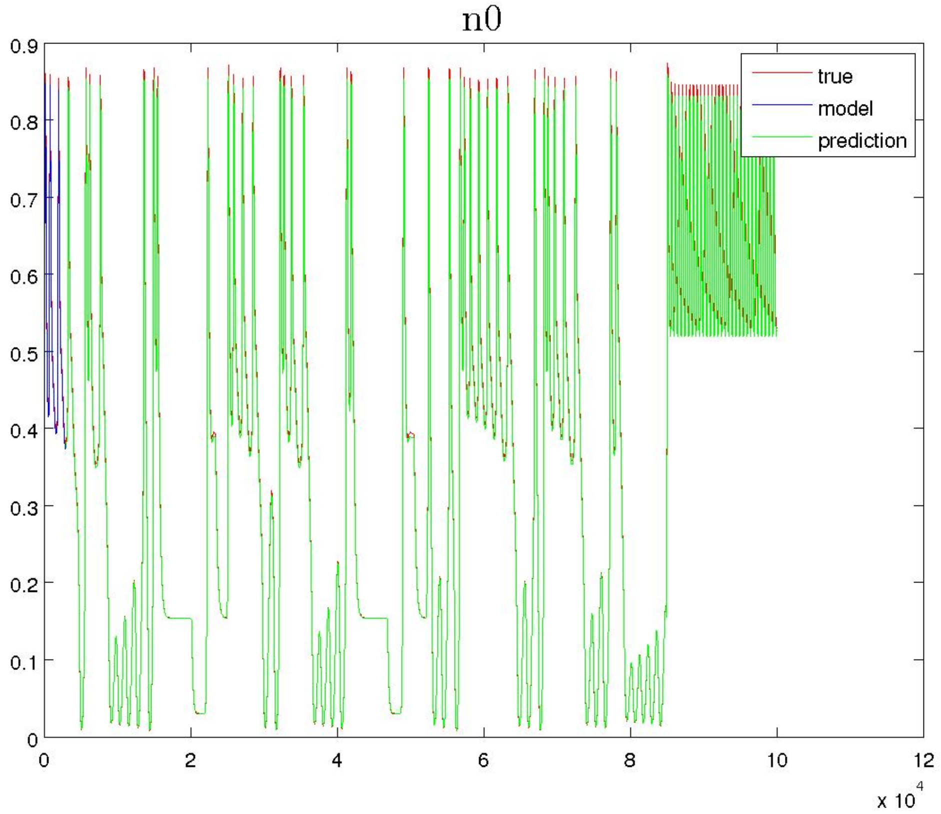

FIG. 5. Results for the *unobserved* gating variables of the potassium channel. The blue lines represent the estimated gating variables n during the assimilation window (3000 data points), whereas the prediction and the true data are represented by the green and red line, respectively. Also in this case the predictions are excellent indicating the validity of the data assimilation methods.