Biophysical Journal
Article

Biophysical Society

Cell Fate Forecasting: A Data-Assimilation Approach to Predict Epithelial-Mesenchymal Transition

Mario J. Mendez,[1,2] Matthew J. Hoffman,[3] Elizabeth M. Cherry,[3,4] Christopher A. Lemmon,[2] and Seth H. Weinberg[1,2,5,*]

[1]Department of Biomedical Engineering, The Ohio State University, Columbus, Ohio; [2]Department of Biomedical Engineering, Virginia Commonwealth University, Richmond, Virginia; [3]School of Mathematical Sciences, Rochester Institute of Technology, Rochester, New York; [4]School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia; and [5]The Dorothy M. Davis Heart and Lung Research Institute, The Ohio State University Wexner Medical Center, Columbus, Ohio

ABSTRACT Epithelial-mesenchymal transition (EMT) is a fundamental biological process that plays a central role in embryonic development, tissue regeneration, and cancer metastasis. Transforming growth factor-β (TGFβ) is a potent inducer of this cellular transition, which is composed of transitions from an epithelial state to intermediate or partial EMT state(s) to a mesenchymal state. Using computational models to predict cell state transitions in a specific experiment is inherently difficult for reasons including model parameter uncertainty and error associated with experimental observations. In this study, we demonstrate that a data-assimilation approach using an ensemble Kalman filter, which combines limited noisy observations with predictions from a computational model of TGFβ-induced EMT, can reconstruct the cell state and predict the timing of state transitions. We used our approach in proof-of-concept "synthetic" in silico experiments, in which experimental observations were produced from a known computational model with the addition of noise. We mimic parameter uncertainty in in vitro experiments by incorporating model error that shifts the TGFβ doses associated with the state transitions and reproduces experimentally observed variability in cell state by either shifting a single parameter or generating "populations" of model parameters. We performed synthetic experiments for a wide range of TGFβ doses, investigating different cell steady-state conditions, and conducted parameter studies varying properties of the data-assimilation approach including the time interval between observations and incorporating multiplicative inflation, a technique to compensate for underestimation of the model uncertainty and mitigate the influence of model error. We find that cell state can be successfully reconstructed and the future cell state predicted in synthetic experiments, even in the setting of model error, when experimental observations are performed at a sufficiently short time interval and incorporate multiplicative inflation. Our study demonstrates the feasibility and utility of a data-assimilation approach to forecasting the fate of cells undergoing EMT.

SIGNIFICANCE Epithelial-mesenchymal transition is a biological process in which an epithelial cell loses epithelial-like characteristics, including tight cell-to-cell adhesion, and gains mesenchymal-like characteristics, including enhanced cell motility. Epithelial-mesenchymal transition is a multistep process in which the cell transitions from the epithelial state to partial or intermediate state(s) to a mesenchymal state. In this study, we use data assimilation to improve prediction of these state transitions. Data assimilation is a technique in which observations are iteratively combined with predictions from a dynamical model to provide an improved estimation of both observed and unobserved system states. We show that data assimilation can reconstruct cell state and predict state transitions using noisy observations while minimizing the error produced by the uncertainty of the dynamical model.

INTRODUCTION

Epithelial-mesenchymal transition (EMT) is a fundamental biological process that plays a central role in embryonic development, tissue regeneration, and cancer metastasis (1–3). The main characteristic of EMT is the transdifferentiation of an epithelial cell to a mesenchymal cell, which

Submitted October 11, 2019, and accepted for publication February 11, 2020.

*Correspondence: weinberg.147@osu.edu

Editor: Kevin Janes.

https://doi.org/10.1016/j.bpj.2020.02.011

includes losing epithelial-type cell-cell adhesion and gaining mesenchymal-type enhanced cell motility. Although EMT is highly controlled and reversible in embryonic development and wound healing, it is often misregulated in a wide array of disease states, including cancer and fibrotic diseases of the liver, kidney, and heart (reviewed in (1)). In disease states, EMT often progresses unchecked, as opposed to embryonic development, in which the process terminates when development is complete. Transforming growth factor-β (TGFβ) is a major and potent inducer of this cellular transition (4–7). Classically, TGFβ-induced EMT has been viewed as an all-or-none switch; however, recent work has demonstrated the existence of multiple steady states in the system, with intermediate or partial states that retain some characteristics of the primary epithelial state but also show features of the mesenchymal state (8–12). Recent evidence further suggests multiple intermediate states exist (13–15). Thus, we can consider TGFβ-induced EMT as comprised of transitions from an epithelial state (E) to an intermediate or partial EMT state or states (P), then to a mesenchymal state (M). Importantly, evidence demonstrates that cells undergoing going EMT can exhibit significant plasticity, capable of being driven in both directions through the EMT program (16–18).

What drives this switch from physiological to pathological EMT? Although many of the pathways that drive EMT are understood, the ability to predict when EMT will occur in a reversible process versus when it will proceed unchecked is difficult. The state transition dynamics of the TGFβ-induced EMT core regulatory pathway is governed by the interactions of a series of transcription factors, including SNAIL1/2 and ZEB1/2, and their respective inhibition mediated by microRNAs miR-34 and miR-200 (19). SNAIL1/2 and ZEB1/2 induce the state transitions of the epithelial cell by promoting the production of the mesenchymal state marker N-cadherin while also decreasing the expression of the epithelial state marker E-cadherin. These transcription factors and miRNAs are linked by several feedback loops, including double-negative feedback loops between SNAIL1 and miR-34—in which SNAIL1 represses the expression of miR-34, which in turn negatively regulates the translation of SNAIL1—and between ZEB and its inhibitor miR-200.

Computational modeling of complex cell signaling pathways has become an established tool to understand signaling mechanisms and make predictions of cell states. However, computational models have several key limitations: even the most detailed biophysical models reproduce the dynamics of a subset of the actual processes occurring in a physiological system. Further, parameters in a computational model are often compiled and extrapolated from a wide range of experimental settings and conditions, and in most cases, parameter values are chosen from experimental mean or median values. Although simulations are often valuable tools for understanding mechanisms and interactions between multiple processes with feedback, in general, it is difficult to perform computational predictions that correspond with a specific individual experiment. That is, simulations may be representative of the "typical" system behavior but not reflective of an individual experiment. Further, long-term computational predictions will often greatly deviate from the truth because of even a small degree of uncertainty in parameters and the highly nonlinear nature of biological systems. Although there have been efforts to use computational models to generate so-called "populations" of simulations to reproduce intertrial experimental variability (20–22), such approaches are generally performed after the experiments to match specific key experimental measurements and not performed in real time as measurements are made.

Experimental measurements of EMT additionally face the technical challenges associated with direct measurement of multiple epithelial and mesenchymal cell markers in the same cell. Although there have been many studies generating significant transcriptomic level data during EMT (23–25), including a series of time points after TGFβ treatment (26), these measurements are generally obtained from a large population of cells and not made in living cells and thus inherently represent a fixed snapshot of the transcriptome for a heterogeneous population. It is not feasible to directly measure all critical EMT-associated cell markers in a given experiment in an individual living cell, resulting in incomplete information on the cell state. Although several cell markers can be quantified with fluorescence intensity measures in individual living cells, this information is insufficient to either fully determine the current cell state or predict future cell state transitions. One limitation in particular is that for the cell markers that are measured in a given experiment, calibrating fluorescence intensity measurements and calculating the corresponding expression levels or concentrations of the epithelial- and mesenchymal-associated cell markers is generally not feasible in real time. Ratiometric measurements are one approach to address these calibration issues. Recently, a stable dual-reporter fluorescent sensor was designed to monitor and mirror the dynamic changes of two key EMT regulatory factors (27), specifically the expression of transcription factor ZEB (quantified indirectly by monitoring a decrease in miR200 binding to a construct containing green fluorescent protein) and epithelial state marker E-cadherin (quantified by E-cadherin promoter-driven expression of red fluorescent protein). Importantly, the dual-reporter sensor enables living measurement of the E-cadherin/ZEB ratio. In this study, we demonstrate that we can incorporate these ratiometric measurements (accounting for experimental noise) into a computational approach known as data assimilation, a well-established technique in the atmospheric science field, to accurately forecast cell fate, including predicting unmeasured cell marker expression levels and, additionally, the timing of EMT-associated state transitions and final cell state.

Data assimilation uses a Bayesian statistical modeling approach to combine high-resolution but imperfect dynamical model predictions with sparse and noisy but repeated experimental observations (28). More specifically, data assimilation is an iterative algorithm in which a previous state estimate (referred to as the background) is updated based on new observations to produce an improved state estimate (referred to as the analysis), which is the maximal likelihood estimate of the model state. The improved state estimate is then used to produce the initial condition for the dynamical model to predict or forecast the future system state estimate, and the process iteratively repeats. An important aspect of this technique is that not all state variables must be measured for this estimation, and furthermore, that unmeasured state variables can also be estimated in time, which is feasible because the dynamics of all state variables are coupled, and this coupling enables prediction of the evolution of unmeasured state variables.

Although data-assimilation approaches have been well utilized in weather forecasting and atmospheric science (28–31), there are relatively few applications in the biomedical sciences (32–42). In this study, we present a data-assimilation approach to reconstruct cell marker expression and predict the timing of the EMT-associated state transitions. We utilize an ensemble Kalman filter (EnKF), which combines limited noisy observations (i.e., only measurement of the E-cadherin/ZEB ratio) with predictions from a computational model of TGFβ-induced EMT (19), to reconstruct the full experimental system and predict the timing of state transitions. We test our approach in proof-of-concept "synthetic" or in silico experiments, in which experimental observations are produced from a known computational model with the addition of noise. We mimic parameter uncertainty in in vitro experiments by incorporating model error that shifts the TGFβ doses associated with the state transitions by either shifting a single parameter or generating a "population" of model parameter sets. We find that EMT-associated dynamics can be successfully reconstructed in synthetic experiments, even in the setting of model error, when experimental observations are performed at a sufficiently short time interval. Furthermore, accurate state reconstruction benefits from incorporating multiplicative inflation, a technique to compensate for underestimation of the true background uncertainty (described further below), which helps manage the influence of model error. Finally, we demonstrate an example in which data-assimilation reconstruction facilitates the accurate long-term prediction of the cell state response to system perturbations that alter cell state. In summary, our study demonstrates an experimentally feasible data-assimilation approach to cell fate forecasting.

METHODS

The main components of the data-assimilation process used in this study are the dynamical systems model (the Tian et al. model, described below), the assimilation algorithm (the EnKF), and observations. Here, to establish the validity and accuracy of our approach in an experimental setting, we use synthetic observations, in which the dynamical system is used to generate a known "truth," with the addition of measurement noise, which can be used for comparison with the data-assimilation state estimate. Data assimilation proceeds iteratively, in which simulations of the dynamical model generate a prediction or forecast, after which the EnKF incorporates observations to generate an improved state estimate, known as the analysis step. The improved state estimate then provides the initial conditions of the next forecast (Fig. 1).

Computational model of EMT

We use the model from Tian and colleagues to represent the core regulatory network of TGFβ-induced EMT, given in Eq. 1 (19). The dynamics of the system are regulated by two coupled or cascading bistable switches, one reversible and the other irreversible. The two bistable switches are regulated by double-negative feedback loops, governing the production of transcription factors SNAIL1/2 and ZEB1/2, respectively, and the inhibition mediated by microRNA miR-34 and miR-200, respectively (Fig. 1 *A*). Model initial conditions and parameters are given in Tables S1 and S2, respectively.

$$\frac{d[\mathrm{T}]}{dt} = k_{0,T} + \frac{k_T}{1 + \left(\frac{[\mathrm{R200}]}{J_T}\right)^{n_{r200}}} - k_{d,T}[\mathrm{T}], \tag{1a}$$

$$\frac{d[\mathrm{s}]}{dt} = k_{0,s} + k_s \frac{\left(\frac{[\mathrm{T}]+[\mathrm{T}]_e}{J_s}\right)^{n_t}}{1 + \left(\frac{[\mathrm{T}]+[\mathrm{T}]_e}{J_s}\right)^{n_t}} - k_{d,s}[\mathrm{s}] \tag{1b}$$

$$\frac{d[\mathrm{S}]}{dt} = \frac{k_S[\mathrm{s}]}{1 + \left(\frac{[\mathrm{R34}]}{J_S}\right)^{n_{r34}}} - k_{d,S}[\mathrm{S}], \tag{1c}$$

$$\frac{d[\mathrm{R34}]}{dt} = k_{0,34} + \frac{k_{34}}{1 + \left(\frac{[\mathrm{S}]}{J_{1,34}}\right)^{n_s} + \left(\frac{[\mathrm{Z}]}{J_{2,34}}\right)^{n_z}} - k_{d,34}[\mathrm{R34}], \tag{1d}$$

$$\frac{d[\mathrm{z}]}{dt} = k_{0,z} + k_z \frac{\left(\frac{[\mathrm{S}]}{J_z}\right)^{n_s}}{1 + \left(\frac{[\mathrm{S}]}{J_z}\right)^{n_s}} - k_{d,z}[\mathrm{z}], \tag{1e}$$

$$\frac{d[\mathrm{Z}]}{dt} = \frac{k_Z[\mathrm{z}]}{1 + \left(\frac{[\mathrm{R200}]}{J_Z}\right)^{n_{r200}}} - k_{d,Z}[\mathrm{Z}], \tag{1f}$$

$$\frac{d[\mathrm{R200}]}{dt} = k_{0,200} + \frac{k_{200}}{1 + \left(\frac{[\mathrm{S}]}{J_{1,200}}\right)^{n_s} + \left(\frac{[\mathrm{Z}]}{J_{2,200}}\right)^{n_z}} - k_{d,200}[\mathrm{R200}], \tag{1g}$$

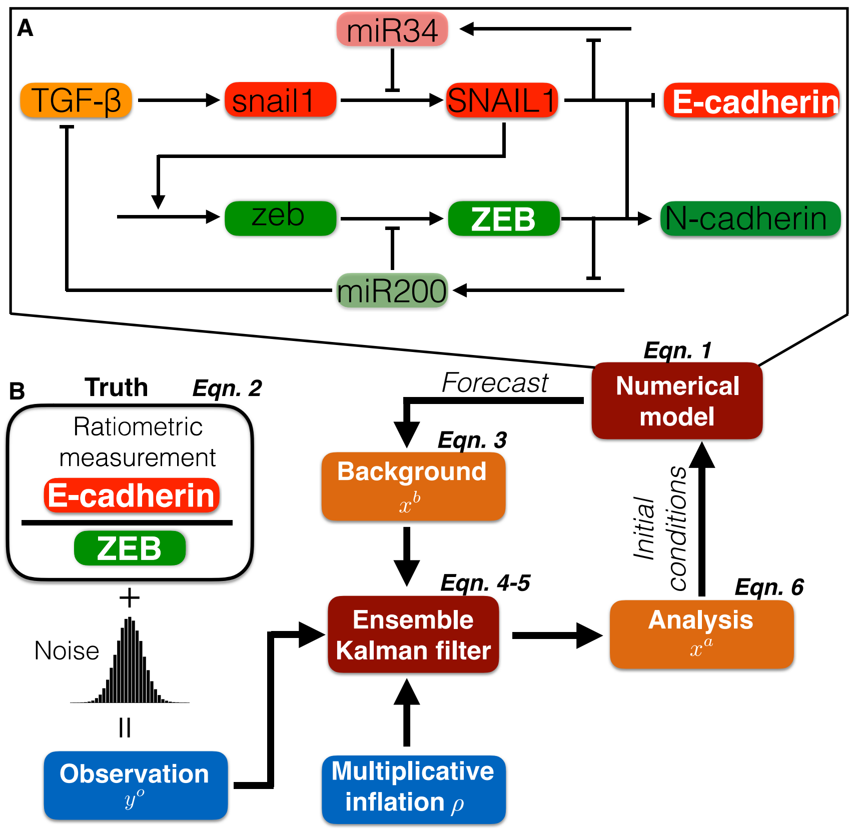

FIGURE 1 Illustration of EMT regulatory network and data assimilation method. (*A*) An illustration of the core regulatory network governing EMT dynamics, modified from Tian et al. (19) and described in the main text, is given. (*B*) A diagram of the data-assimilation method is shown: synthetic observations are generated from ratiometric measurements of E-cadherin/ZEB from the "truth" system, plus the addition of Gaussian noise. The numerical model (in *A*) generates ensembles of forecasts. Combining the forecasts and observations, the EnKF yields the maximal likelihood estimator for the system state (the analysis), which provides initial conditions for the next iteration. To see this figure in color, go online.

$$\frac{d[\mathrm{E}]}{dt} = \frac{k_{e,1}}{1+\left(\frac{[\mathrm{S}]}{J_{e,1}}\right)^{n_s}} + \frac{k_{e,2}}{1+\left(\frac{[\mathrm{Z}]}{J_{e,2}}\right)^{n_z}} - k_{d,e}[\mathrm{E}], \tag{1h}$$

and

$$\frac{d[\mathrm{N}]}{dt} = \frac{k_{n,1}}{1+\left(\frac{[\mathrm{S}]}{J_{n,1}}\right)^{n_s}} + \frac{k_{n,2}}{1+\left(\frac{[\mathrm{Z}]}{J_{n,2}}\right)^{n_z}} - k_{d,n}[\mathrm{N}]. \tag{1i}$$

Exogenous TGFβ ($[\mathrm{T}]_e$) increases the production of snail1 messenger RNA (mRNA) ([s]), activating the first double-negative feedback loop by increasing the translation of SNAIL1 protein ([S]), which in turn inhibits production of miR-34 ([R34]), the inhibitor of SNAIL1 translation. SNAIL1 activates the second double-negative feedback loop by increasing the production of zeb mRNA ([z]), increasing translation of ZEB protein ([Z]), which in turn inhibits production of miR-200 ([R200]), the inhibitor of ZEB translation. Both SNAIL1 and ZEB suppress the epithelial state marker E-cadherin ([E]) and promote the mesenchymal state marker N-cadherin ([N]). Suppression of miR-200 production further removes inhibition of endogenous TGFβ, a positive feedback that promotes the first feedback loop and results in an irreversible phenotype switch.

A representative simulation demonstrating the transition from an epithelial to mesenchymal state is shown in Fig. 2 *A*. The initial conditions are defined consistent with an epithelial state, i.e., high E-cadherin and low N-cadherin expression. A constant dose of 3 μM exogenous TGFβ is applied for 20 days. An initial increase in SNAIL1 is associated with a moderate decrease in E-cadherin and increase in N-cadherin expression. The simulation illustrates the existence of a state with intermediate levels of both E-cadherin and N-cadherin expression, which is defined as a partial EMT state. A secondary increase of SNAIL1 promotes a subsequent production of ZEB and, in turn, production of endogenous TGFβ. The transition from a partial EMT to a mesenchymal state is associated with a further decrease in E-cadherin and increase in N-cadherin expression.

Motivated by the recent development of a novel dual-reporter sensor for the EMT state (27), which emits fluorescence proportional to E-cadherin and ZEB, we also illustrate the dynamics of the ratio between E-cadherin/ZEB, which exhibits a decrease ranging several orders of magnitude. As described below, this ratiometric measurement will serve as the observations used in our data-assimilation approach, demonstrating the utility of this metric that can be measured experimentally.

In Fig. 2 *B*, we illustrate the model responses to varying exogenous TGFβ doses for initial conditions in the epithelial (*blue*), partial (*red*), or mesenchymal (*green*) states. We plot N-cadherin expression at the end of a 20-day time interval. For initial conditions in the epithelial state, increasing exogenous TGFβ results in a step-like increase in the final N-cadherin expression level, with an intermediate level corresponding with a partial EMT state and the elevated level corresponding with the mesenchymal state. Interestingly, for initial conditions in a partial EMT state, hysteresis is observed such that the TGFβ doses associated with the epithelial-to-partial state (E-P) transition and the partial-to-mesenchymal state (P-M) transition depend on the initial state. Further, for an initial mesenchymal state, the irreversibility of the second bistable switch results in the maintenance of the mesenchymal state for all TGFβ doses, even in the absence of any exogenous TGFβ added.

Data assimilation

Data-assimilation methods are a class of algorithms that are used to improve the state estimation and forecasting ability of dynamical systems

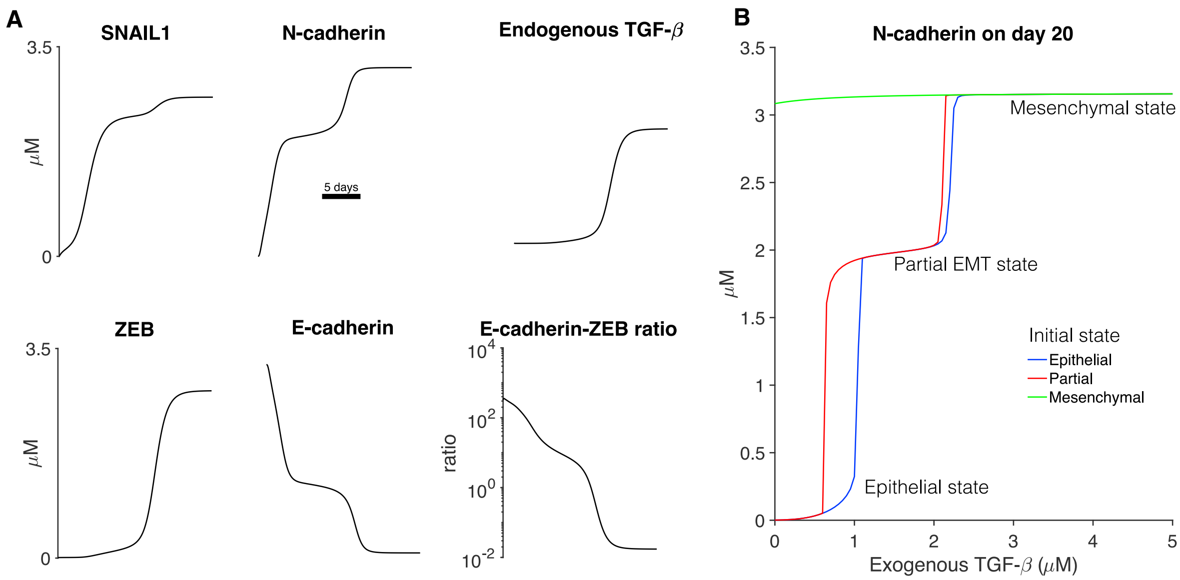

FIGURE 2 TGFβ induces EMT via a partial or intermediate EMT state transition. (*A*) The time course of key epithelial and mesenchymal markers or a ratio of markers is shown as a function of time after the addition of exogenous TGFβ. (*B*) The expression level of N-cadherin on day 20 is shown as a function of the exogenous TGFβ dose for different initial conditions. The step-like response illustrates distinct cell states, corresponding with epithelial, partial or intermediate EMT, and mesenchymal states. Parameters: (*A*) exogenous TGFβ = 3 μM. To see this figure in color, go online.

by combining observations of the system with a numerical model of the system dynamics. Much of the research on data assimilation for large systems originates in the atmospheric science community (30,43–45), in which it is a crucial piece of numerical weather prediction. Although data assimilation was originally designed to improve forecasts by improving the current state estimate—and therefore delaying some of the chaotic drift of the forecast—data assimilation has also been used to estimate and correct parameters of the forecast model. Beyond the Earth's atmosphere, data assimilation has been used on the Martian atmosphere, as well as on oceans, estuaries, lakes, and biological systems (32,38,39,46,47).

EnKF

In this work, data assimilation is completed using an EnKF, which is an extension of the linear Kalman filter for nonlinear problems. The EnKF attempts to estimate the most likely state of the system given a prior estimate of the state, a (potentially sparse) set of observations of the system, and uncertainty estimates for both the state and the observations. The EnKF is an iterative process illustrated in Fig. 1 *B* and detailed below.

For this problem, the state space at time t is a column vector of the nine model variables at this time,

$$x^t(t) = ([\mathrm{T}](t), [\mathrm{s}](t), [\mathrm{S}](t), [\mathrm{R34}](t), [\mathrm{z}](t), [\mathrm{Z}](t),$$

$$[\mathrm{R200}](t), [\mathrm{E}](t), [\mathrm{N}](t))^T. \quad (2)$$

The prior state estimate, which here initially comes from a previous model run, is called the background state and is denoted x^b. Estimating the uncertainty in the background—denoted $\mathbf{P}^b$—is typically the most difficult aspect of the approach, especially because this uncertainty is state dependent. In an EnKF, the background uncertainty is assumed to be Gaussian, and the mean and covariance are parameterized by a small number of model states. This is similar to a Monte Carlo approach but with fewer ensemble members (typically on the order of 10–100) than would be needed to fully sample the space.

The algorithm used here is an ensemble transform Kalman filter that is the local ensemble transform Kalman filter algorithm without the localization (28). Following the notation of Hunt et al. (28), given a set of background states $x^{b(i)}$, the background is computed as the mean of the ensemble members,

$$x^b = \frac{1}{k}\sum_{i=1}^{k} x^{b(i)}, \quad (3)$$

where k is the ensemble size, and the covariance is given by the ensemble sample covariance,

$$\boldsymbol{P}^b = \frac{1}{k-1}\sum_{i=1}^{k}\left(x^{b(i)} - x^b\right)\left(x^{b(i)} - x^b\right)^T \quad (4)$$

The Kalman filter finds the state that minimizes the cost function

$$J(\tilde{x}) = \left(\tilde{x} - x^b\right)^T\left(\boldsymbol{P}^b\right)^{-1}\left(\tilde{x} - x^b\right) + \left[y^o - H(\tilde{x})\right]^T\boldsymbol{R}^{-1}\left[y^o - H(\tilde{x})\right], \quad (5)$$

where y^o is the vector of observations, $\mathbf{R}$ is the covariance of these observations, and H is a map from the model space to the observations space (which is typically lower dimensional). The state that minimizes the cost function in the subspace spanned by the ensemble members is called the analysis and is denoted x^a. The analysis error covariance matrix in ensemble space, $\tilde{\mathbf{P}}^a$, can be computed in ensemble space as $\tilde{\boldsymbol{P}}^a = [\rho^{-1}(k-1)\boldsymbol{I} + \boldsymbol{Y}^{b^T}\boldsymbol{R}^{-1}\boldsymbol{Y}^b]^{-1}$. Here, ρ is a multiplicative inflation parameter. Multiplicative inflation is a way of compensating for the fact that the small ensemble size tends to lead to underestimation of the true background uncertainty. Multiplying the covariance matrix by a constant greater than 1 (ρ here) is the simplest and most computationally efficient way of correcting for this underestimation. The inflation factor ρ is a tunable parameter for the assimilation. The columns of the $\mathbf{Y}^b$ matrix are the perturbations of the background ensemble members mapped into observation space. Mathematically, the jth column of $\mathbf{Y}^b$ is $y_j^b = H(x^{b(j)}) - y^b$, where $y^b = \frac{1}{k}\sum_{j=1}^{k} H(x^{b(j)})$ is the mean of the background ensemble in observation space.

The analysis covariance is then used to transform the background ensemble perturbations into analysis ensemble perturbations according to $\boldsymbol{X}^a = \boldsymbol{X}^b[(k-1)\tilde{\boldsymbol{P}}^a]^{1/2}$. Finally, the new analysis mean is computed as

$$\boldsymbol{x}^a = \boldsymbol{x}^b + \boldsymbol{X}^b \tilde{\boldsymbol{P}}^a \boldsymbol{Y}^{b^T} \boldsymbol{R}^{-1}\left(\boldsymbol{y}^o - \boldsymbol{y}^b\right). \tag{6}$$

The analysis mean is added to each column of $\boldsymbol{X}^a$ to generate the analysis ensemble members. The analysis ensemble members then become initial conditions for the next forecast (i.e., numerical integration of Eq. 1), which generates the background members x^b that in turn are used for the next analysis time. Numerical integration is performed in MATLAB (The MathWorks, Natick, MA) using the ode15s ordinary differential equation solver. Descriptions of the variables in the EnKF method are provided in Table 1. A more detailed description of the algorithm, including derivations, can be found in (28). In this study, we consider ensemble sizes k between 5 and 50, multiplicative inflation factors ρ between 1 and 1.6, and observation intervals (i.e., intervals between analysis steps) Δt_{obs} between 2 and 48 h.

Numerical experiments

For a given data-assimilation trial, the truth system was initialized with all state variables in the epithelial state. To initialize each ensemble member of the background, a separate model simulation was performed with a random TGFβ dose (uniformly sampled between the 0 and given dose for that trial) for a random duration (uniformly sampled between 0 and 20 days), and final state variable concentrations were chosen for the ensemble initial state. Synthetic observations were generated from the truth system using a ratiometric measurement of E-cadherin and ZEB. Observational measurement noise or error was reproduced by adding to the true ratio a Gaussian random variable with a mean of 0 and standard deviation equal to 10% of the true ratio magnitude. Minimal E-cadherin and ZEB concentrations were set to 1.1×10^{-5} μM to avoid negative or undefined ratio values.

We assess the accuracy of a given data-assimilation trial with two approaches: 1) we calculate the root mean-square deviation (RMSD) between the true system and the average of the analysis ensembles, summing over all state variables, as a function of time:

$$RMSD(t) = \sqrt{\frac{1}{m}\sum_{j=1}^{m}\left(x_j^a(t) - x_j^t(t)\right)^2}, \tag{7}$$

where $x_j^a(t)$ and $x_j^t(t)$ are the jth element of the analysis and truth m-dimensional vectors, respectively, at time t. We calculate the area under the RMSD versus the time curve to quantify error for a single trial; 2) for each ensemble, we predict the timing of the E-P and, when appropriate, P-M state transitions and compare it with the true timing of these transitions. These calculations are performed as follows: after each analysis step, each ensemble is simulated for the remaining time of the 20-day simulation duration. The E-P and P-M state transitions are determined as the time when N-cadherin expression increases above 1.5 and 3.0 μM, respectively. Finally, we average the predicted thresholds over all ensembles. This calculation is repeated for each analysis step. Related to this second calculation, in a subset of data-assimilation trials, we also calculate the accuracy of the prediction of the final cell state (i.e., epithelial, partial, or mesenchymal) at the end of the simulation.

We first consider the case in which the same parameters are used to simulate both the truth and ensembles, using the baseline parameter set in Tian et al. (19). To assess the data-assimilation approach in the context of parameter uncertainty, we then also consider the influence of model error by two approaches: 1) modifying a single model parameter, specifically increasing the snail1 mRNA degradation rate $k_{d,s}$ from 0.09 to 0.108, which alters the dynamics of the first double-negative feedback loop and shifts the TGFβ doses associated with the E-P and P-M state transitions to higher levels (see Fig. 5); and 2) randomly scaling a large subset of model parameters (28 out of 44) to generate a "population" of model parameter sets, which alters the TGFβ dose and time dependence of EMT dynamics and qualitatively reproduces variability observed in in vitro experiments (see Fig. 9).

Both in the presence and absence of model error, we assessed RMSD and state transition predictions for varying data-assimilation properties. Specifically, we varied the time interval between observations/analysis steps Δt_{obs}, the number of ensembles k, and multiplicative inflation ρ. In summary analysis, for each set of data-assimilation properties, measures were averaged over 25 trials to account for randomness in the initialization process. For statistical analysis, the Kolmogorov-Smirnov test was used to assess distribution normality. When appropriate, Student's t-tests were performed to compare distribution means. The Wilcoxon signed rank test was used to compare RMSD error for trials with data assimilation with the error for a trial without data assimilation. The Mann-Whitney U-test and Kruskal-Wallis test were used to assess statistical significance between trials with different data-assimilation properties.

TABLE 1 DA Variables

Notation	Description
k	number of ensemble members
m	model space dimension
l	number of observations
ε	Gaussian random variable added to the truth to form observations
ρ	multiplicative inflation factor
Δt_{obs}	observation interval
H	map from model space to observation space
x^t	m-dimensional true state vector
$x^{b(i)}$	m-dimensional vector of background ensemble member i
$y^{b(i)} = H(x^{b(i)})$	l-dimensional vector of the background state estimate mapped to observation space
x^a	m-dimensional analysis vector
$x^b = \frac{1}{k}\sum_{i=1}^{k} x^{b(i)}$	m-dimensional background state estimate vector
$y^b = \frac{1}{k}\sum_{i=1}^{k} y^{b(i)}$	l-dimensional vector of the mean of $y^{b(i)}$
$y^o = H(x^t) + \varepsilon$	l-dimensional observations vector
$\boldsymbol{X}^b$	$m \times k$ matrix of background ensemble member perturbations from their mean x^b
$\boldsymbol{Y}^b$	$l \times k$ matrix of background ensemble perturbations in observation space from their mean y^b
$\boldsymbol{X}^a$	$m \times k$ matrix of analysis ensemble member perturbations from their mean x^a
$\boldsymbol{P}^b$	$k \times k$ ensemble sample covariance
$\boldsymbol{R}$	$l \times l$ observation covariance matrix
$\tilde{\boldsymbol{P}}^a$	$k \times k$ analysis error covariance matrix

Notation and description of key variables defined and utilized in the EnKF method for DA. See text for details.

RESULTS

Data assimilation in the absence of model error

A representative data-assimilation experiment is shown in Fig. 3, for which the truth (*black line*) and ensembles (*dashed blue lines*) utilize the same model parameters

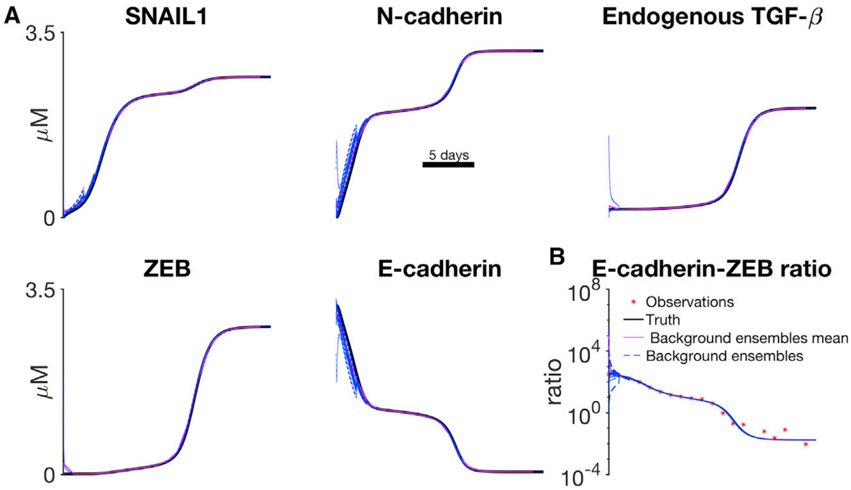

FIGURE 3 Data assimilation reconstructs unobserved EMT dynamics. (*A*) The truth (*black*), ensembles (*blue*), and ensemble mean (*magenta*) are shown as a function of time for key epithelial and mesenchymal markers. (*B*) Observations (*red stars*) of the E-cadherin/ZEB ratio are shown for each observation interval Δt_{obs}. Parameters: exogenous TGFβ = 3 μM. Observation interval Δt_{obs} = 24 h, number of ensembles k = 10, multiplicative inflation ρ = 1. To see this figure in color, go online.

(i.e., no model error) and using synthetic E-cadherin/ZEB ratiometric observations (*red stars*) with an observation interval of 24 h, 10 ensembles, and no multiplicative inflation (i.e., $\rho = 1$). In both truth and ensemble simulations, 3 μM TGFβ is applied at time 0. The background ensemble mean (*magenta line*) followed the true E-cadherin/ZEB ratio within the initial 48 h (i.e., two analysis steps; Fig. 3 *B*). Importantly, the dynamics of the unobserved state variables, including SNAIL1, ZEB, E-cadherin, N-cadherin, and endogenous TGFβ, were also reconstructed successfully by the background ensemble mean after 48 h (Fig. 3 *A*).

We first quantified the accuracy of the data-assimilation experiments by measuring RMSD error relative to the true system (Fig. 4 *A*). RMSD error with data assimilation (*blue line*) demonstrates small increases near the timing of state transitions; however, the RMSD error is greatly reduced compared with trials without data assimilation (*magenta*). We next quantified the accuracy of the data-assimilation corrected simulations to predict the true timing of state transitions from the E-P state and P-M state. These predictions were performed as follows: after each analysis step, each ensemble was simulated for the remainder of

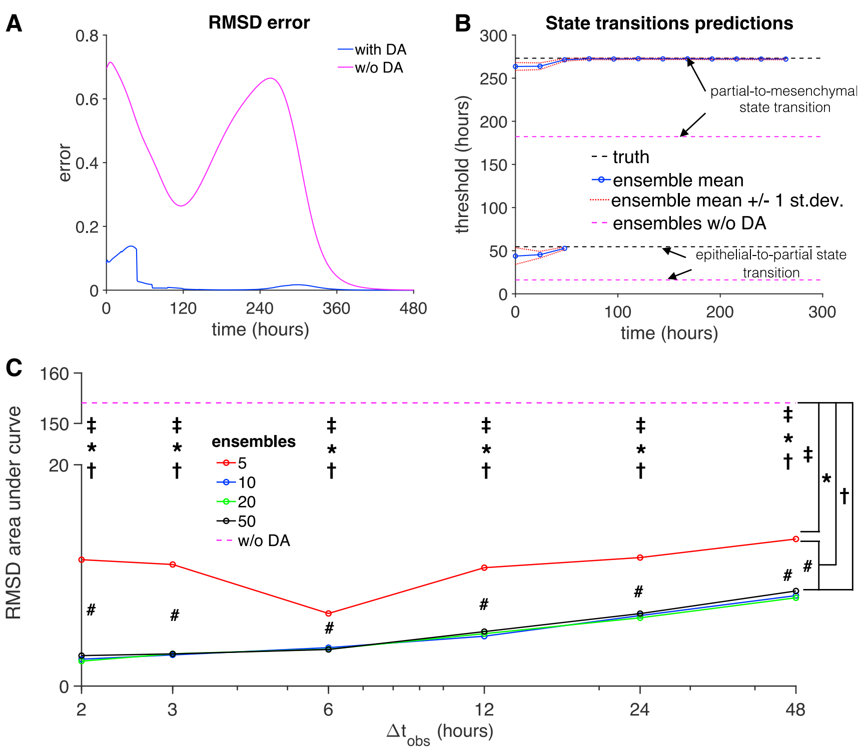

FIGURE 4 Data-assimilation error and predictions. (*A*) The RMSD error (Eq. 7) for a trial with (*blue*) and without (*magenta*) data assimilation (DA) is shown as a function of time. The trial without DA was initialized randomly, but no analysis steps were performed. (*B*) The truth (*dashed black*) value and ensemble mean (*blue*) and ensembles without DA mean (*dashed magenta*) predictions for the E-P state transition and P-M state transition are shown as a function of time. Ensemble mean predictions $\pm$ 1 standard deviation (*dashed red*) are also shown. Parameters for (*A* and *B*) are the same as Fig. 3. (*C*) The average value for the area under the RMSD versus time curve is shown as a function of the observation interval Δt_{obs} for different number of ensembles k (*solid lines*) and without DA (*dashed magenta*). In (*C*), 25 trials per DA condition: Wilcoxon signed rank test, †$p < 10^{-10}$ (DA, k = 10, 20, 50 (grouped) vs. without DA), *$p < 10^{-10}$ (DA, k = 5, 10, 20, 50 (grouped) vs. without DA), ‡$p < 10^{-10}$ (DA, k = 5 vs. without DA); Mann-Whitney U-test: #$p < 10^{-10}$ (DA, k = 10, 20 50 (grouped) vs. DA, k = 5). Comparisons were performed for each Δt_{obs} value. To see this figure in color, go online.

the 20-day duration, and the timing of each transition was determined (if the transition was predicted). We then averaged transition threshold over all ensembles. Because we perform this prediction after each analysis step, we thus report the predicted threshold as a function of time (Fig. 4 *B*). We find that the ensemble mean predictions (*solid blue lines*) initially underestimate the timing of both the E-P and P-M state transitions, which occurs because a subset of ensembles are initialized in or near a partial state. However, the data-assimilation predictions converge toward the true timing of E-P and P-M state transitions (*black dashed lines*) within 48 h (i.e., two analysis steps), which, importantly, is before either transition occurs (see Fig. 4 *B*). In contrast, without data-assimilation corrections, predictions of both transitions are underestimated (*dashed magenta lines*).

We next varied the observation interval Δt_{obs} and number of ensemble members k (Fig. 4 *C*). For each condition, we calculated the area under the RMSD curve, averaging over 25 trials to account for randomness in the initialization process. Consistent with Fig. 4 *A*, for all observation intervals and ensemble sizes, RMSD error area was significantly less than the error for trials without data assimilation (*dashed magenta*). We found that observation interval was a highly significant factor ($p < 10^{-10}$), whereas the effect of the ensemble size was near significance ($p = 0.0758$) because of the difference in RMSD error area between $k = 5$ ensemble member trials and other ensemble sizes. Beyond $k = 5$ ensemble members, we find that varying ensemble size had minimal effect on the RMSD error area, i.e., the effect of ensemble size was not significant ($p = 0.524$), whereas the effect of the observation interval remained highly significant ($p < 10^{-10}$). Indeed, we find that RMSD error area increased approximately linearly as the observation interval increased, i.e., larger error for fewer observations and analysis steps. These results demonstrate that in the absence of model error, this data-assimilation approach can greatly reduce error in predictions of system variables and state transition timing in a manner that depends on the interval for observations while minimally depending on ensemble size.

Data assimilation in the presence of model error

We next consider several conditions in which model error is introduced and further consider key factors that determine the predictive power of the data-assimilation approach. We consider two different situations: 1) model error is introduced by altering one parameter, which results in different steady-state behavior for a given TGFβ dose; and 2) model error is introduced by altering a large subset of all parameters (here, 28 parameters) to generate a "population" of parameter sets.

Model error due to inaccuracy in one parameter

For the first situation, we consider a modified parameter set in which the snail1 mRNA degradation rate $k_{d,s}$ is increased from 0.09 to 0.108. This modification (*red line*, Fig. 5 *A*) alters the dynamics of the first double-negative feedback loop and right shifts the TGFβ doses associated with the E-P and P-M state transitions to higher levels relative to the baseline parameter set (*black line*). Specifically, we consider four exogenous TGFβ doses (*vertical dashed blue lines*) that result in four combinations of final steady states between systems with the baseline parameter set and the modified parameter set: 1) 1.1429 μM, which produces a partial EMT state for the baseline system and an epithelial state for the modified system; 2) 1.675 μM, which produces a partial EMT state for both systems but with altered dynamics; 3) 2.626 μM, which produces a mesenchymal state on the baseline system and a partial EMT state for the modified system; and 4) 4.05 μM, which produces a mesenchymal state for both systems but with altered dynamics (Fig. 5 *B*).

For the next series of synthetic experiments, the true system uses the baseline parameter set, whereas the ensemble simulations forecast using the modified parameter set with an increased $k_{d,s}$. Fig. 6 illustrates the performance of the data-assimilation algorithm for $k = 20$, $\rho = 1$, and

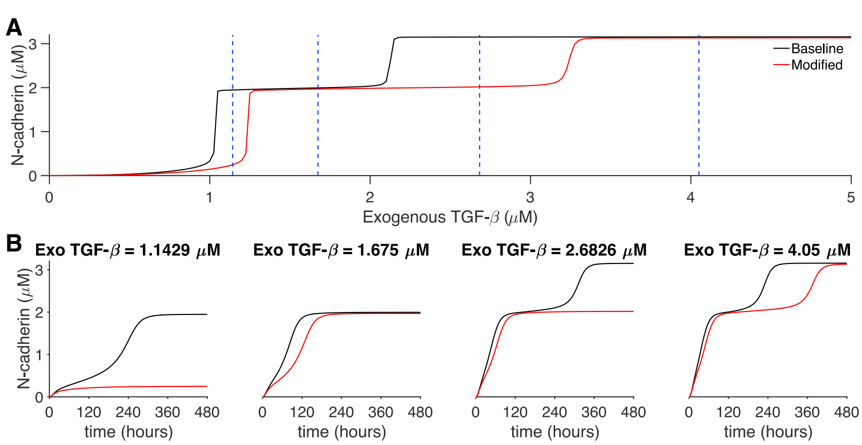

FIGURE 5 Modified EMT dynamics and TGFβ dose dependence. (*A*) N-cadherin expression level on day 20 is shown as a function of the exogenous TGFβ dose for baseline (*black*) and modified (*red*) parameter sets. (*B*) N-cadherin expression as a function of time is shown for the four TGFβ doses denoted in (*A*) (*vertical dashed blue lines*). Parameters: $k_{d,s} = 0.09$ (baseline), $k_{d,s} = 0.108$ (modified). Other model parameters are unchanged. To see this figure in color, go online.

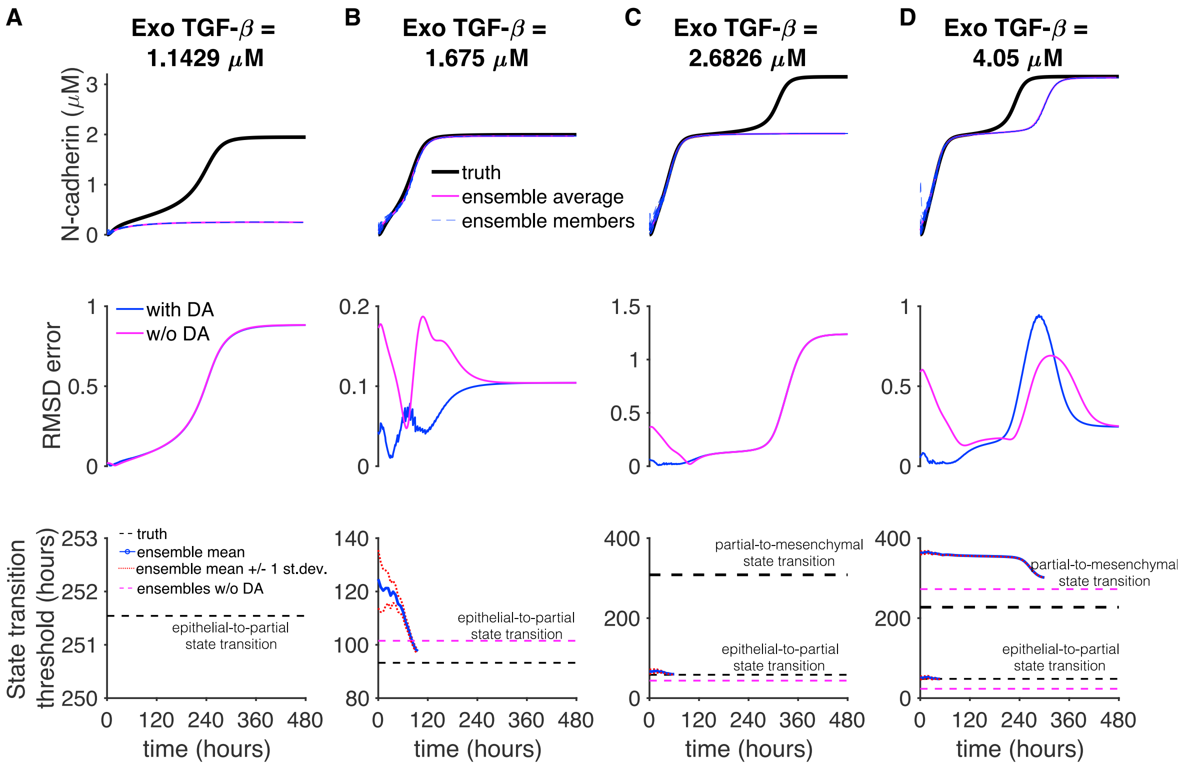

FIGURE 6 DA fails to reconstruct EMT dynamics with model error, with frequent observations and without multiplicative inflation. (*Top*) N-cadherin expression for the truth (*black*), ensemble members (*dashed blue*), and ensemble mean (*magenta*); (*middle*) RMSD error with (*blue*) and without (*magenta*) DA; and (*bottom*) the truth value (*dashed black*) and ensemble mean (*blue*) and ensembles without DA mean (*dashed magenta*) predictions for state transitions are shown as a function of time for (*A–D*) four TGFβ doses. Parameters: observation interval $\Delta t_{obs} = 6$ h, number of ensembles $k = 20$, multiplicative inflation $\rho = 1$. Truth system: $k_{d,s} = 0.09$ (baseline), ensembles: $k_{d,s} = 0.108$ (modified). To see this figure in color, go online.

$\Delta t_{obs} = 6$ h. For exogenous TGFβ of 1.1429 μM (dose 1), data assimilation failed to reconstruct the true system dynamics (Fig. 6 *A*, *top panel*) because the ensemble mean remained in the epithelial state, whereas the true system transitioned to a partial EMT state. RMSD error was comparable to simulations without data assimilation (*middle panel*), and furthermore, the E-P transition of the true system was not predicted at any point throughout the simulation (*bottom panel*). For exogenous TGFβ of 1.675 μM (dose 2), data assimilation successfully reconstructed the true system dynamics (Fig. 6 *B*, *top panel*), with a reduction of the RMSD error before the E-P transition (middle panel). The ensemble prediction of the E-P transition was initially overestimated (consistent with the modified parameter set; see Fig. 5); the prediction improved throughout the simulation but only accurately predicted the timing immediately preceding the transition (*bottom panel*).

For exogenous TGFβ of 2.626 μM (dose 3), data assimilation failed to reconstruct the true system dynamics because the ensemble mean remained in a partial EMT state, whereas the true system transitioned to a mesenchymal state (Fig. 6 *C*, *top panel*), similar to the first example. Similarly, P-M transition of the true system was not predicted at any point throughout the simulation, although the E-P transition was accurately predicted (*bottom panel*). Finally, for exogenous TGFβ of 4.05 μM (dose 4), data assimilation successfully reconstructed the dynamics of the E-P transition of the true system; the ensemble mean also reproduced the P-M transition, although at a later time than the true system (Fig. 6 *D*). The ensemble predictions of both the E-P and P-M transition were initially overestimated; the E-P transition prediction converged on the true timing, whereas the P-M prediction improved but did not converge before the transition occurred in the true system. Thus, in general, these data-assimilation conditions failed to predict the timing of state transitions and only predicted steady-state response when the steady state of true and modified parameter sets were the same (i.e., simulations without data assimilation would also predict the steady-state response).

We next consider the effect of incorporating multiplicative inflation by increasing ρ to 1.4 (Fig. 7). For all TGFβ doses, the ensemble mean accurately reproduces the dynamics of the true system (*top panels*), and the RMSD error remained lower than simulations without data assimilation (*middle panels*). Furthermore, for all exogenous doses, the predictions of the state transitions timings converged to the true values (*bottom panels*). Importantly, although for TGFβ doses 1 and 3, the E-P and P-M transitions, respectively, are initially not predicted to occur, after a sufficient time, these transitions are predicted and indeed converge

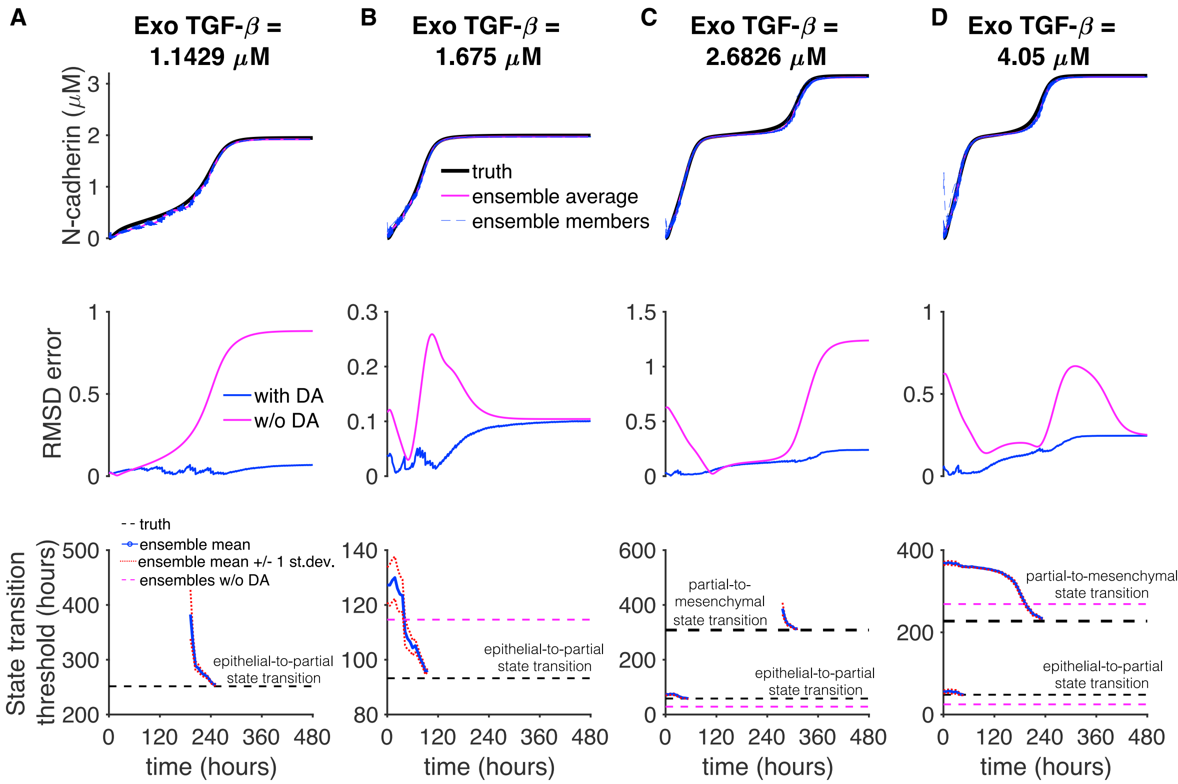

FIGURE 7 DA successfully reconstructs EMT dynamics with model error, with frequent observations and multiplicative inflation. (*Top*) N-cadherin expression for the truth (*black*), ensemble members (*dashed blue*), and ensemble mean (*magenta*); (*middle*) RMSD error with (*blue*) and without (*magenta*) DA; and (*bottom*) the truth value (*dashed black*) and ensemble mean (*blue*) and ensembles without DA mean (*dashed magenta*) predictions for state transitions are shown as a function of time for (*A–D*) four TGFβ doses. Parameters: observation interval $\Delta t_{obs} = 6$ h, number of ensembles $k = 20$, multiplicative inflation $\rho = 1.4$. Truth system: $k_{d,s} = 0.09$ (baseline), ensembles: $k_{d,s} = 0.108$ (modified). To see this figure in color, go online.

to the true value (Fig. 7, *A* and *C*, *bottom panels*). Thus, we find that incorporating multiplicative inflation greatly improves the predictions of state transitions sufficiently before their respective occurrence in the presence of model error. We further investigate how the inclusion of multiplicative inflation improves state reconstruction in the presence of model error by plotting the unobserved state variables for the four TGFβ doses (Figs. S1–S4). We find that the model error in the snail1 mRNA degradation rate results in reduced snail1 mRNA levels, relative to the truth, both with and without multiplicative inflation, as would be expected. However, in the presence of multiplicative inflation, microRNAs miR-34 and miR-200 are altered, reducing their levels closer to the truth and resulting in more accurate predictions of SNAIL1 and ZEB expression levels, respectively, and in turn more accurate prediction of E-cadherin and N-cadherin levels and associated EMT state. Kadakia and colleagues found similar compensatory shifts in the setting of model error using data assimilation to estimate model parameters in neurons (48).

In the Supporting Material, using the data-assimilation parameters in Fig. 7 with $\rho = 1.4$ and $\Delta t_{obs} = 6$ h, we consider the case for which the true system used the modified parameter set and the ensembles used the baseline parameters (Fig. S5). Similar to Fig. 7, we find that the ensemble mean successfully reproduces the dynamics of the true system (*top panels*), and RMSD error is consistently less than simulations without data assimilation (*middle panels*). We also investigate the importance of the observation interval by increasing Δt_{obs} to 24 h (Fig. S6). We find that in general, in contrast to Fig. 7, even with the inclusion of multiplicative inflation, the data-assimilation approach fails to reproduces the dynamics of the true system: for TGFβ dose 1, the steady-state dynamics are predicted but the timing of the E-P transition is not, whereas for dose 3, the steady-state dynamics are not predicted, and the P-M transition is not predicted to occur at any point during the simulation. Thus, we find that even with the inclusion of multiplicative inflation, infrequent observations and analysis step corrections can result in a failure to predict the true system dynamics and associated state transitions.

These simulations suggest that data assimilation with properly determined parameters can accurately reproduce the true system dynamics for conditions in which model error without data assimilation would lead to a failure to predict a state transition (as in Figs. 6 and 7) or to an erroneous prediction of a state transition (as in Fig. S5). To quantitatively summarize the predictive power of the data-assimilation approach with the presence of model error, we next performed a parameter study over a wide range of TGFβ doses and different data-assimilation properties, varying k, ρ, and Δt_{obs}. We performed 25 trials for each case and quantified the mean RMSD error area under curve over these trials (Fig. 8). For nearly all conditions, RMSD error area was significantly less with data assimilation compared with trials without data assimilation. Across all TGFβ doses, ensemble size was a marginally significant factor on the RMSD error area, whereas observation interval and multiplicative inflation were highly significant (see Fig. 8). We find that in the absence of multiplicative inflation ($\rho = 1$), RMSD error area is only slightly better than simulations without data assimilation (*magenta line*) for most TGFβ doses and generally did not depend on observation interval or ensemble size. Consistent with Figs. 6, 7, and S6, incorporating multiplicative inflation reduced error, generally more so for sufficiently small observation intervals (typically less than 24 or 48 h). For intermediate multiplicative inflation $\rho = 1.2$, error decreased as the observation interval decreased. For larger multiplicative inflation ρ of 1.4 or 1.6, error had a U-shaped dependence, with a minimal error for Δt_{obs} of 6 h typically, for all TGFβ doses. This demonstrates that although observations are necessary with a sufficiently frequent interval to reduce error, too-frequent observations (and analysis steps) results in overcorrection (i.e., ensemble collapse) and an increase in error. Further, for larger multiplicative inflation ($\rho = 1.4$–1.6) and short observation intervals Δt_{obs} below 6 h, a subset of data-assimilation trials became unstable because of state variables in the nonphysiological regime, resulting in a dramatic increase in the mean RMSD error, which occurred more frequently for larger ensemble sizes.

We additionally performed the same broad data-assimilation parameter study, for which the true system used the modified parameter set and the ensemble background used the baseline parameter set (as in Fig. S5). Similar to the previous study (Fig. 8), without multiplicative inflation, RMSD error generally does not depend on the ensemble size or observation interval, although for this case, it is generally lower than simulations without data assimilation (Fig. S7). Also, as in Fig. 8, error decreases for smaller observation interval Δt_{obs} for moderate multiplicative inflation ($\rho = 1.2$),

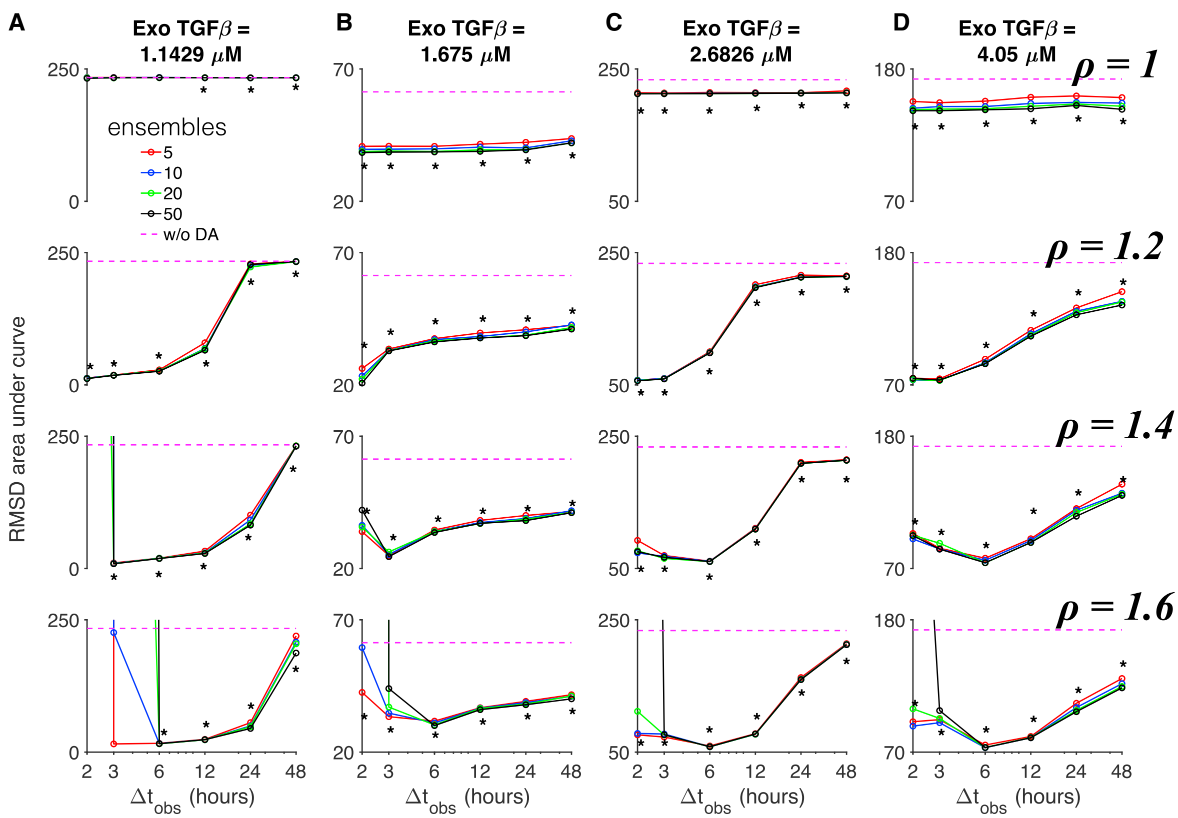

FIGURE 8 Optimal observation intervals and multiplicative inflation reduce prediction error of EMT dynamics, with model error. The average value for the area under the RMSD versus time curve is shown as a function of the observation interval Δt_{obs} for different number of ensembles k (*solid lines*) and without DA (*dashed magenta*) for (*A–D*) four TGFβ doses and multiplicative inflation values ρ (different *rows*). Parameters: truth system, $k_{d,s} = 0.09$ (baseline); ensembles: $k_{d,s} = 0.108$ (modified). 25 trials per DA condition: Wilcoxon signed rank tests, $*p < 0.05$ (DA, $k = 5, 10, 20, 50$ (grouped) vs. without DA). Kruskal-Wallis tests for k: $p = 0.0698, 0.00013, 0.0159, 0.0334$ (TGFβ doses 1–4, respectively); for Δt_{obs}: $p < 10^{-10}$ (all TGFβ doses); for ρ: $p < 10^{-10}$ (all TGFβ doses). Note that trial variability tended to decrease for larger Δt_{obs} such that statistical differences between DA and without DA trials were found, even when the average values were similar (as in *A*, $\rho = 1$). To see this figure in color, go online.

whereas error generally has a U-shaped dependence for larger multiplicative inflation ($\rho = 1.4$–1.6), with a minimum near Δt_{obs} of 3 or 6 h. We similarly find dramatic increases in error for larger ρ and small Δt_{obs}. Thus, across a wide range of data-assimilation experiments incorporating model error and multiple TGFβ doses resulting in different EMT states, we find that moderate multiplicative inflation and short observation intervals consistently demonstrate the smallest predictive error.

Model error in a large population of parameter sets

Finally, we investigate the predictive accuracy of the data-assimilation approach in the setting of model error associated with a large "population" of model parameter sets. Specifically, the truth system was drawn from a population of model parameter sets that reproduces the biological variability in cell state that is observed in experiments, in which multiple cell states (i.e., epithelial, partial, or mesenchymal state) are observed at a given TGFβ dose concentration and duration (8). To generate a population of model parameter sets, random scaling factors for the basal and regulated production, transcription, and translation rates and degradation rates for endogenous TGFβ, snail1 mRNA, SNAIL1, miR-34, zeb mRNA, ZEB, and miR-200 (i.e., the first 28 parameters in Table S2) were chosen from a log-normal distribution with median 1, following an approach similar to Sobie (20). Distribution parameter $\sigma = 0.075$, the standard deviation of the distribution of the log-transformed variables, determines the parameter variability. As illustrated in Fig. 9, we drew 1000 random parameter sets and simulated the response to five different exogenous TGFβ doses. The time course for N-cadherin expression and the distribution of different cell states are shown in Fig. 9, *A* and *B*, respectively. This population of simulations qualitatively reproduces the cell state distribution observed by Zhang and colleagues (cf. Fig. 6 *C* in (8)). Steady-state distributions of cell state further illustrate the heterogeneous mixture of cell state for moderate TGFβ doses (Fig. S8). Note that we increase the rate constant scaling factor $\kappa = 2$ in these and subsequent simulations to more closely match the time course of the experimental results.

To quantify the accuracy of the data-assimilation approach in the setting of physiological model error, we performed simulations with the truth system using each individual parameter set from the population and the ensemble background system using the baseline parameter set. The ensemble mean value of the N-cadherin expression level was used to predict the current cell state, and the ensembles were simulated for the remaining time in the 15-day simulation to predict the final cell state. The

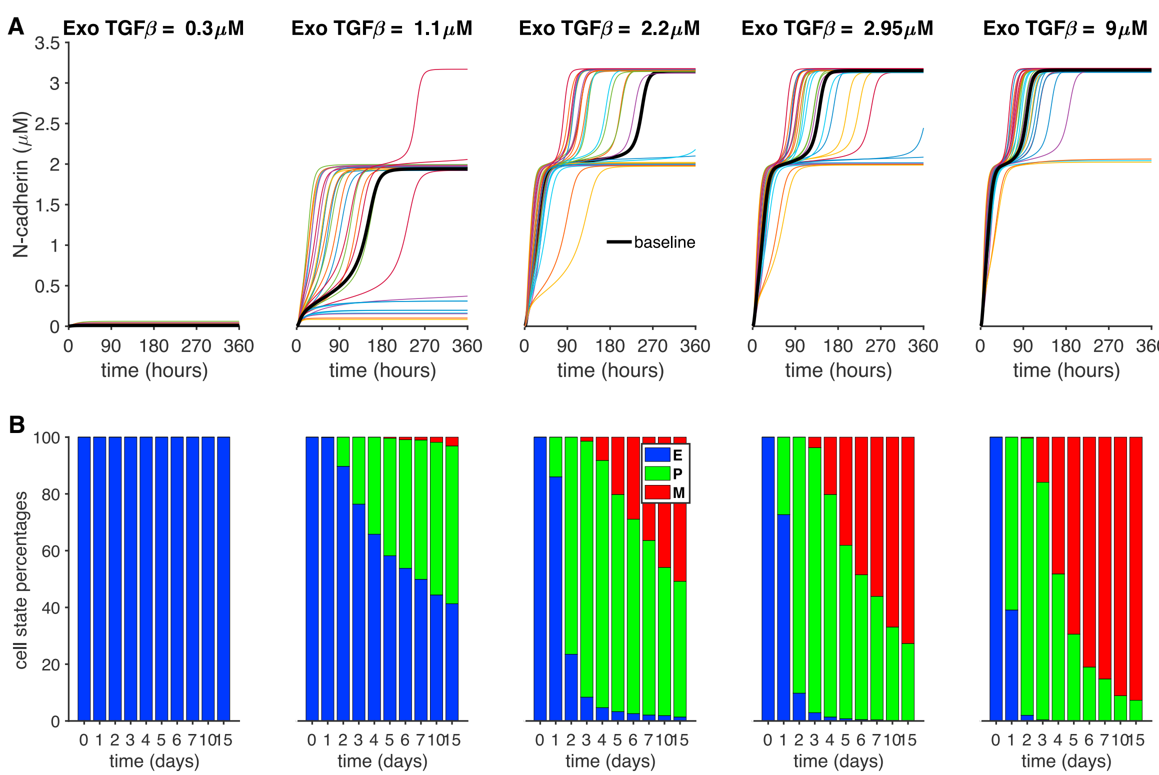

FIGURE 9 Population of model parameter sets reproduces experimental heterogeneity in cellular state. (*A*) N-cadherin expression for the baseline parameter set (*thick black line*) and a subset of the population members is shown as a function of time for different exogenous TGFβ doses. Note that only 50 simulations out of 1000 are shown for clarity. (*B*) The percentage of cells in the epithelial (*blue*), partial (*green*), and mesenchymal (*red*) states is shown as a function of time for different exogenous TGFβ doses. Parameters: baseline parameter values in Table S2, except $\kappa = 2$. Population parameter sets were randomly drawn as described in the text. To see this figure in color, go online.

accuracy of both current and final state predictions was calculated over all 1000 parameter sets and as a function of time.

In Fig. 10, we consider data assimilation with multiplicative inflation for several exogenous TGFβ doses. For a low exogenous TGFβ dose (0.3 μM, *first column*), all population parameter sets and the baseline parameter set remain in the epithelial cell state for all 15 days such that current and final cell state predictions are trivially correct for 100% of parameter sets. For an intermediate exogenous TGFβ dose (1.1 μM, second column), most parameter sets result in either a final epithelial or partial EMT state, with a small final mesenchymal cell state subset (see Fig. 9 *B*). Over all 1000 parameter sets, we find that the current state prediction (*solid red*) accuracy remains near 90% for the entire 15-day simulation duration (Fig. 10 *A*). The final cell state prediction is initially just above 50% and then increases near day 7 (*solid blue*).

We further quantify prediction accuracy for subsets of the population based on the true final cell state. For this intermediate TGFβ dose, the baseline parameter set results in a final partial EMT state. Thus, for the final partial EMT state subset, the current state prediction (*solid red*) is near 100%, and the final state prediction (*solid blue*) is 100% for all time points (Fig. 10 *C*, *second column*). For the final epithelial state subset (Fig. 10 *B*, *second column*), current cell state prediction accuracy is initially at 100% bcause the entire population and the baseline parameter set are both initially in the epithelial state. Current cell state prediction accuracy decreases with time because the baseline parameter set does not accurately reconstruct all of population subset that remain in the epithelial state; however, predictive accuracy is still quite high, near 90% for all time points. The final cell state prediction is initially 0%; however, predictive accuracy increases sharply above 50% at day 7. Similarly, for the final mesenchymal state subset (Fig. 10 *D*, *second column*), current cell state predictions are near 90% for all time points, although fluctuating because of the small size of this subset. The final cell state prediction is also initially 0% and then increases around day 5. Importantly, for both the epithelial and mesenchymal final state subset (i.e., parameter sets in which the true final state differs from the baseline parameter set), the final state is accurately predicted in a majority of parameter sets by around day 10, several days before the end of the 15-day simulation.

For a larger exogenous TGFβ dose (2.2 μM, *third column*), most parameter sets result in either a final partial

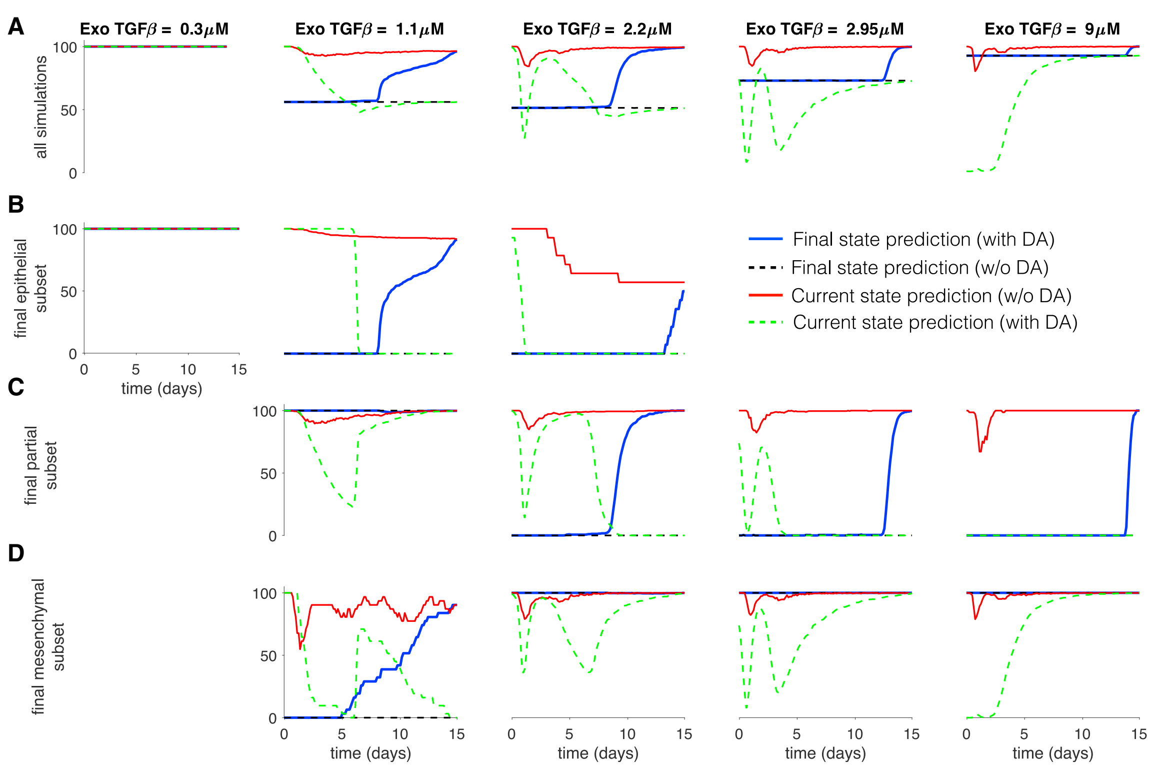

FIGURE 10 Accuracy of state prediction in a population of model parameter sets, with model error and multiplicative inflation. The accuracy of the DA approach prediction of the current and final (at day 15) cell state (i.e., epithelial, partial, or mesenchymal) is shown as a function of time (*solid red* and *blue lines*, respectively). Corresponding predictions without DA are shown in dashed green and black lines, respectively. The accuracy out of all 1000 simulations is shown in (*A*), and the subsets of simulations in which the final cell state is epithelial, partial, and mesenchymal are shown in (*B*)–(*D*), respectively. Parameters: $\Delta t_{obs} = 3$ h, $\rho = 1.4$, $k = 20$. Note that the values for the final state prediction without DA (*dashed black*) are, by definition, either 0% or 100% for all time points in (*B*)–(*D*), depending on the final state for the baseline parameter set. To see this figure in color, go online.

or mesenchymal state, with a small final epithelial state subset. For this larger TGFβ dose, the baseline parameter set results in a final mesenchymal state, and thus, the current and final state prediction accuracy for the final mesenchymal subset is near and at 100%, respectively (Fig. 10 *D*, *third column*). The final state prediction is initially 0% for the final epithelial subset (Fig. 10 *B*) and final partial subset (Fig. 10 *C*) and then increases around days 12 and 8, respectively. For higher exogenous TGFβ doses (2.95 and 9 μM, *fourth* and *fifth columns*), most parameter sets result in a final mesenchymal state such that predictions of the final mesenchymal subset are highly accurate. Final cell state predictions of the partial subset are initially inaccurate and increase sharply just before the end of the 15-day simulation. Thus, we find that the data-assimilation approach can accurately predict the current cell state with high accuracy at all time points. Further, the final cell state is accurately predicted for a wide range of TGFβ dose conditions in the majority of parameter sets in the population. Importantly, for cases in which the true final state differs from that of the baseline parameter set, the final state can still be accurately predicted in a majority of simulations several days before the end of the experiment. In the Supporting Material, we illustrate that final cell states are not accurately predicted in the absence of multiplicative inflation (Fig. S9), and final state prediction accuracy improves with multiplicative inflation and shorter observation intervals (Fig. S10).

For a final analysis, we investigate the differences between the distribution of the parameter sets in the population for which the final state is accurately or inaccurately predicted. Specifically, we consider the TGFβ dose of 1.1 μM (Fig. 10, *second column*), and we identify the parameter values in sets that correspond with accurate prediction of the final cell state at day 10. In Fig. 11, we plot histograms of all parameter value scaling factors (such that a value of 1 corresponds with the baseline parameter value) for correct (*blue*, 788 values) and incorrect (*red*, 212 values) final state predictions. Interestingly, although we find differences in the distributions and distribution means (*vertical dashed lines*) in several of the parameters, the largest statistical differences occur for several of the degradation rate and Hill constant parameters (k_d- and J-values, respectively). This suggests that model error due to inaccurate degradation rates or Hill constants may be particularly critical in the prediction of cell state using this data-assimilation approach, although a more thorough study of the relationship between model error, cell state prediction,

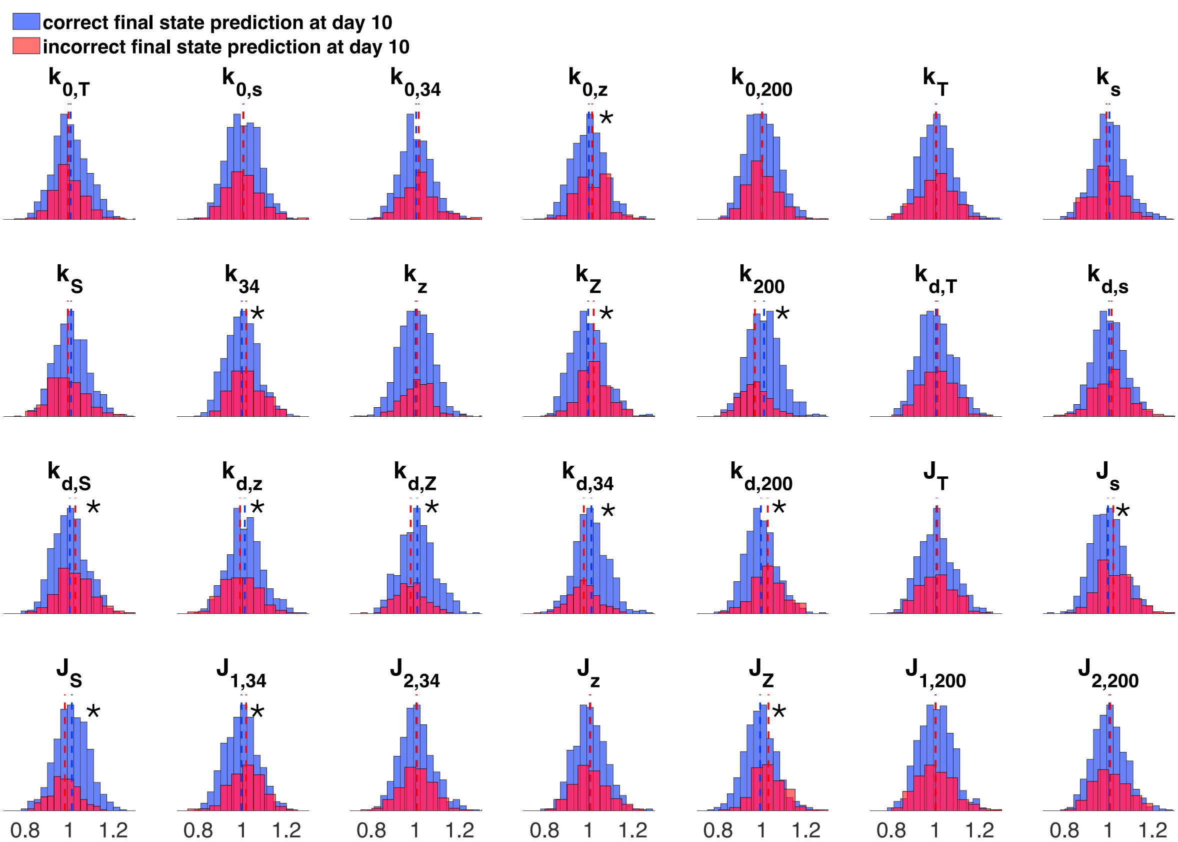

FIGURE 11 Parameter distributions for correct and incorrect final state predictions. Histograms of parameter scaling factors for correct (*blue*) and incorrect (*red*) predictions of the final cell state after 10 days are shown for each model parameter. Scaling factor of 1 corresponds with the baseline parameter value. Vertical dashed lines denote distribution means. Parameters: exogenous TGFβ = 1.1 μM, Δt_{obs} = 3 h, ρ = 1.4, k = 20. Student's t-test: *$p < 0.01$ (correct versus incorrect prediction parameter scaling factors). To see this figure in color, go online.

and data-assimilation properties is a focus of ongoing work.

DISCUSSION

Summary of main findings

In this study, we use a data-assimilation approach for a series of synthetic experiments to forecast cell fate in the setting of EMT. First, proof-of-concept in silico experiments are performed in which experimental observations are produced from a known computational model with the addition of noise but in the setting of no model error, i.e., all model parameters are assumed to be known. In the absence of model error, EMT dynamics are successfully reconstructed, generally within 48 h of observations.

To mimic parameter uncertainty present in in vitro experiments, we introduce model error in a manner that shifts the TGFβ doses and dynamics associated with state transitions because of discrepancies in either a single or many model parameters. In the presence of model error, EMT dynamics are successfully reconstructed using the data-assimilation approach, incorporating multiplicative inflation and an optimal observation interval. That is, sufficiently frequent observations are needed to observe and predict EMT transitions, whereas a sufficient interval between observations and the addition of multiplicative inflation mitigate overconfidence in model predictions. The inclusion of multiplicative inflation increases relative confidence in observations and overcomes the incorrect model dynamics driven by model error, in particular by altering the levels of (unobserved) regulatory microRNAs. With these ideal conditions, even in the presence of model error, the timing of EMT state transitions and the final cell state behavior are successfully predicted. Further, we find that these results negligibly depend on the number of ensembles in the EnKF, demonstrating that a computationally efficient approach using fewer ensembles is feasible and sufficient.

EMT is a process characterized by a phenotypic shift in epithelial cells to motile and oftentimes invasive mesenchymal cells. This tightly regulated process is fundamental in the generation of new tissues and organs during embryogenesis and is a key factor in tissue remodeling and wound healing (1–3). Although EMT is critical for development, its misregulation is implicated in many diseases, including cardiac fibrosis, cirrhosis, and cancer. In these disease states, it is not only crucial to understand what drives EMT to better understand the pathology, it is equally important to predict the timing of EMT-associated state transitions, with an eye toward developing effective therapies (i.e., system perturbations) to reverse EMT-related disorders. One of the main complications with making such predictions is the limited number of EMT-associated markers that can be observed experimentally in an individual live cell experiment; all experimental measurements are inherently providing an incomplete snapshot of the system state at a given moment in time.

The data-assimilation approach presented in this study demonstrates several key advances in the prediction of EMT dynamics: 1) expression levels of unmeasured EMT-associated cell markers are accurately reconstructed and predicted based on a single ratiometric measurement of two cell markers. Although these predictions are inherently limited by the details of the biophysical model from which they are based, this approach can be easily adapted to utilize more detailed predictive models of cell signaling to predict expression levels of additional unmeasured markers; 2) by integrating a predictive biophysical model with experimental observations, we can accurately predict future events, specifically the timing of cell phenotype state transitions and final cell state. This technique can be more generally applied as a tool to probe responses to various experimental perturbations applied at different stages and timings throughout the EMT process, such as changes in TGFβ dose or agonists and antagonists of different signaling pathways, that can be predicted and then applied in real time. Both of these extensions are the focus of ongoing future work.

Applications of real-time predictive forecasting

As a preliminary example, we highlight the utility of real-time predictive cell forecasting and demonstrate how it can be used to improve an experimental protocol. As noted above, there is recent evidence to suggest that the partial state transition between the epithelial and mesenchymal state is in fact comprised of multiple intermediate states (13–15), and thus, it would be of interest to develop an experimental protocol to suppress the transition into the mesenchymal state and promote the stability of a partial EMT state to investigate the heterogeneity of different cell markers in this hybrid state. However, as shown by Zhang and colleagues (8) (reproduced in Figs. 9 and S8), application of a moderate exogenous TGFβ dose would be expected to result in a mix of epithelial, partial, and mesenchymal cell states because of variability within a population of cells.

In Fig. 12, we illustrate an example in which the truth system is governed by a parameter set generated by the population approach described above, and the ensemble parameter set is the baseline parameter set. For the applied exogenous TGFβ dose, the baseline parameter set results in a partial EMT state (*gray line*, Fig. 12 *A*), whereas the truth system is in the mesenchymal state after 10 days (*black line*). Using the data-assimilation approach, the ensemble mean (*solid blue*) reconstructs the system for a specified time period (3 days, *vertical dashed line*). At this time point, the ensembles are projected until the simulation end and, without additional observations, accurately forecast the later transition into the mesenchymal state (*dashed blue line*). At the time of this projection, we can apply a system

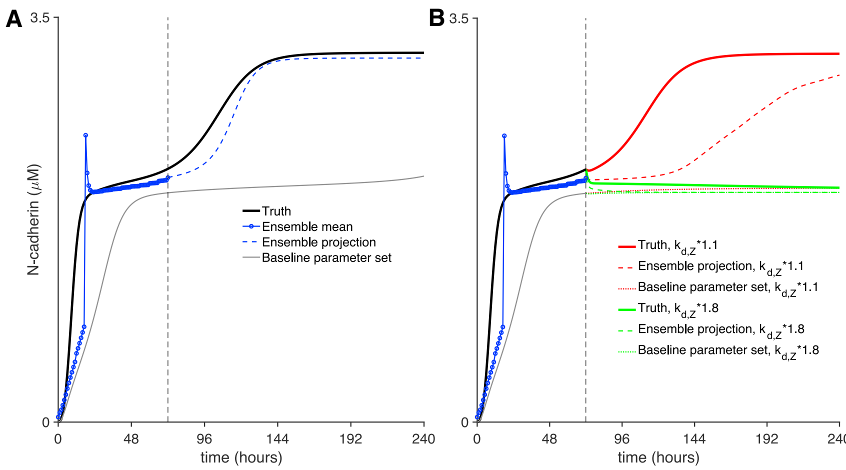

FIGURE 12 Real-time prediction of cell state and perturbations. (*A*) N-cadherin expression for the truth (*solid black*) and baseline parameter set (*solid gray*) is shown as a function of time. The ensemble mean (*solid blue*) from the DA approach is shown as a function of time up to 72 h (*vertical dashed line*). After 72 h, the ensemble projection (the average of all ensemble forecasts) is shown as a function of time for the remainder of the simulation. (*B*) After 72 h, N-cadherin expression for the truth (*thick*), ensemble projection (*dashed*), and baseline parameter set (*thin*) is shown as a function of time for the cases of the ZEB degradation rate $k_{d,Z}$ scaled by a factor of 1.1 (*red lines*) and 1.8 (*green lines*). Parameters: exogenous TGFβ = 2.15 μM, observation interval $\Delta t_{obs} = 6$ h, multiplicative inflation $\rho = 1.4$, number of ensembles $k = 20$. Ensembles: baseline parameter set (Table S2), except $\kappa = 2$, and $k_{d,Z}$ altered as described in the text. To see this figure in color, go online.

perturbation expected to suppress the mesenchymal transition and predict the response. In this example, we consider a hypothetical agonist that increases the ZEB degradation rate $k_{d,Z}$. We find that a small increase does not suppress the mesenchymal transition (*thick red line*), whereas a larger increase maintains a partial EMT state (*thick green line*), and importantly, both of these final cell states are accurately predicted by ensemble projection (*dashed red* and *green lines*, respectively), whereas the baseline parameter predicts a partial EMT state for both cases (*dotted red* and *green lines*, respectively). In future work, we plan to more systematically investigate different perturbations and protocols to determine conditions in which such approaches would be predicted to be successful.

Extensions to complete model error and model selection

Although the synthetic studies performed here consider examples of model error that arise because of inaccurate parameter values, which subsequently shift the TGFβ dose dependence of the EMT state transitions, the fundamental mechanisms governing these transitions are conserved, i.e., both the true and forecasting models are governed by cascading bistable switches. For a final demonstration of the utility of the data-assimilation approach, we consider a more complete example of model error in which the true system is governed by differing molecular mechanisms. We consider the EMT model from Lu and colleagues (49), which models the core dynamics of SNAIL and ZEB signaling with detailed microRNA-mediated regulation and similarly reproduces epithelial, partial EMT, and mesenchymal states. However, in this model formulation, the state transitions are governed by a ternary chimera switch (8), in which the miR-34/SNAIL subsystem is monostable, whereas the miR-200/ZEB subsystem forms a ternary switch, i.e., the subsystem has three steady states that correspond with the epithelial, partial EMT, and mesenchymal states.

In Fig. 13, we consider an example in which the forecasting ensemble model is governed by the baseline Tian et al. model (19), whereas the true system is governed by the Lu et al. (49) model, augmented with Eqs. 1h and 1i governing E-cadherin and N-cadherin dynamics, respectively,

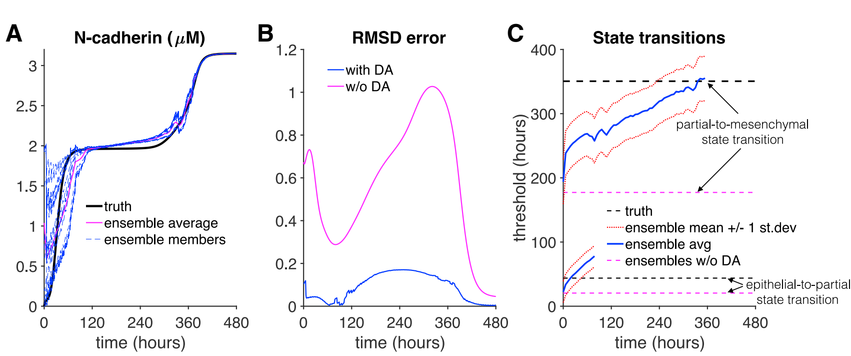

FIGURE 13 DA successfully predicts state transitions in the setting of complete model error. (*A*) N-cadherin expression for the truth (*black*), ensemble members (*dashed blue*), and ensemble mean (*magenta*); (*B*) RMSD error with (*blue*) and without (*magenta*) DA; and (*C*) the truth value (*dashed black*) and ensemble mean (*blue*) and ensembles without DA mean (*dashed magenta*) predictions for state transitions are shown as a function of time. Molecular counts in the Lu et al. model were rescaled assuming comparable maximal concentrations for appropriate comparison with the Tian et al. model. Parameters: observation interval $\Delta t_{obs} = 6$ h, multiplicative inflation $\rho = 1.1$, number of ensembles $k = 20$. True system: equations given by Eq. S4.4 and parameters in Tables S1 and S4 in (49). EMT stimulus (analogous to TGFβ) $I(t) = I_{max}\min\{t/t_{ramp}, 1\}$, I_{max} = 150,000 molecules, t_{ramp} = 96 h. Ensembles: baseline parameter set (Table S2), TGFβ = 3 μM. To see this figure in color, go online.

for comparison with the Tian et al. model. We find that the data-assimilation approach ensemble average (*magenta*) reasonably reconstructs N-cadherin expression levels (Fig. 13 *A*), although individual ensembles appear "jagged" or noisy (*blue*). Further, the timing of the P-M transition is also successfully predicted, whereas the approach overestimates the timing of the E-P transition, likely because of the differences in dynamics governing SNAIL (Fig. 13 *C*). The predicted dynamics of the unobserved state variables are closer to the true system compared with ensembles without data assimilation; however, interestingly, most unobserved state variables still differ from the true system (Fig. S11), demonstrating that complete reconstruction of true system state dynamics is not necessary for accurate prediction of state transitions. Thus, despite the significant differences in governing dynamics in the setting of complete model error, we find that the data-assimilation approach can predict EMT state transitions.

Although further work is needed to broadly characterize the approach in this setting of complete model error, our preliminary assessment found that the predictive accuracy was highly sensitive to the data-assimilation parameters. Interestingly, this suggests that the data-assimilation approach may be utilized to determine the model formulation most representative of the true system, i.e., model selection. By considering a "competition" of candidate models for the forecasting ensemble model, a selected model would ideally be accurate for a wide range of experimental conditions and robust to different data-assimilation algorithmic parameters.

Prior data-assimilation applications to biological systems

A few prior studies have applied data-assimilation approaches to different biological systems. Several studies, including those by two of the authors of this work, have reconstructed excitable cell dynamics for various levels of scale and complexity. Munoz and Otani applied a Kalman filter on single cardiac cells to predict the dynamical behavior of state variables not directly observed such as intracellular ionic concentrations (50). Hoffman and colleagues used an EnKF approach to reconstruct complex electrical rhythms in one-dimensional and three-dimensional cardiac tissues and similarly found that the addition of inflation in the data-assimilation algorithm was pivotal to improve prediction accuracy while also showing minimal influence of ensemble size (38,39). Several studies by Hamilton and colleagues have applied data-assimilation approaches to predict dynamics of neural electrical activity, including determination of neural network connectivity (33) and reconstruction of intracellular ion concentrations (34) and of intracellular potential (35). Moye and Diekman apply two different classes of data-assimilation approaches to improve estimates of both neural cell state and model parameters for different types of bifurcation behavior (40). Ullah and Schiff applied Kalman filters to predict unobserved states in neurons and small neural networks (36,37).

Data assimilation has also successfully been applied to improve predictions of a human brain tumor growth in in silico experiments using synthetic magnetic resonance images (32). Using predictions from a simple tumor growth model and integrating measurements from a more detailed model, the data-assimilation algorithm successfully produced accurate qualitative and quantitative analysis of brain tumor growth. A similar Kalman filter approach has also been applied to dynamical state reconstruction, with a focus on prediction of unobserved state variables and parameter estimation in models of mammalian sleep dynamics (42) and blood glucose levels (41). In several studies, Abarbanel and colleagues used a variational data-assimilation approach to estimate both state variables and parameters in neurons and neuronal networks (48,51,52). In general, prior work has focused on reconstructing physiological system dynamics, often predictions of unobserved system states, with several applications to excitable cells and tissue. Although the dynamics of these systems are often governed by excitable, oscillatory, and bursting behavior, here we consider a system with distinct dynamics that are regulated by multiple bistable switches, and we show that data assimilation can successfully reconstruct cell state dynamics and transitions in such a system that governs cell phenotype.

Limitations

Because this study is an initial proof-of-concept demonstration of using data assimilation to predict EMT dynamics, there are several key limitations to be addressed in future studies. The Tian et al. model used in this study represents the core regulatory pathway of TGFβ-induced EMT. Although the model is based on key experimental findings of the interactions of critical transcription factors and microRNAs regulating the EMT process (19), there are other signaling pathways (e.g., Wnt and β-catenin signaling (53,54)) involved in EMT that are not accounted for. However, our approach can be naturally extended to account for the details of additional signaling pathways. As an initial test, we only consider signaling occurring in a single cell and do not consider spatial interactions occurring within a multicellular tissue during EMT. Predictions within a multicellular tissue are inherently more challenging because multicellular tissues can exhibit spatial heterogeneity in cell phenotype, both initially and as a function of time. Model development of the spatial interactions during the EMT process is complex, and this challenge is indeed an area of ongoing work within our lab (55) and others (18,56–58). As described by Hunt and colleagues (28), the EnKF can be further extended to account for spatial localization and interacting spatial dynamics, and we plan to extend the approach demonstrated here to multicellular tissues in the future as well. We note that although the EnKF is just one of several possible extensions of the linear Kalman filter for nonlinear problems, with other nonlinear approaches including the extended Kalman filter, unscented Kalman filter, and particle filter (59), the EnKF was developed for data assimilation with high-dimensional systems for which the covariance matrices are difficult to compute directly. Data assimilation using a model composed of hundreds of cells undergoing EMT would result in such a high-dimensional system, and thus, this future application to multicellular tissue motivated our choice of the EnKF with the single-cell model.

In weather forecasting, the hourly forecast is typically more accurate than 10-day predictions, and we similarly find that predictions throughout the data-assimilation process generally were more accurate closer to the timing of the predicted state transitions, i.e., short-term predictions were more accurate than long-term predictions. This can be observed in Figs. 4 *B* and 7 (*bottom panels*), as the state transition predictions become more accurate at later time points and converge to the true value, and also in Fig. 10, as the final state predictions (*blue lines*) increased greatly after ~7 days. Indeed, our numerical experiments demonstrate that the timescale for accurate predictions is on the order of 5–10 days, depending on conditions. Whereas short-term predictions can be quite accurate despite model error, long-term predictive accuracy suffers specifically because of parameter inaccuracy because the impact of inaccurate parameters on system dynamics compounds as the time since the most recent observation increases. In this study, we consider model error in the setting of either a single inaccurate parameter or multiple simultaneous inaccurate parameters. As noted above, many data-assimilation studies have focused on the estimation of both state variables and model parameters. The parameter estimation problem is a natural extension of the approach described here by including estimated parameters in the state variable vector with trivial dynamics (i.e., derivative equal to 0). Variational data-assimilation methods such as 4D-Var (60) have been particularly successful at parameter estimation in neuronal and cardiac models (40,48,51,61). One challenge with the inclusion of parameter estimation is the determination of which parameters to estimate. Estimation of all model parameters could result in compensatory changes that may improve current state prediction but not necessarily produce accurate reconstruction of true parameter values and thus limit long-term projections and prediction of the responses to system perturbations. However, importantly, our analysis of differences between parameter sets with accurate and inaccurate final cell state prediction suggested that accurate reconstruction of degradation rates and Hill constants may be critical and thus is a natural target for our initial parameter estimation investigation.

In this study, we demonstrated that the EnKF can accurately predict EMT dynamics in the setting of model error without additionally estimating model parameters. This is particularly useful in the situation in which the inaccurate model parameters are not known. Although the approach did require thorough investigation of data-assimilation parameters—specifically, observation interval and multiplicative inflation—importantly, we were able to identify algorithmic parameters that performed consistently across a wide range of conditions and model parameter regimes. In future work, we plan to perform a systematic study of physiological model error in EMT dynamics with parameter estimation. Finally, our long-term goal is to consider realistic biological model error, that is, using our approach with in vitro observations from fluorescence measurements of the E-cadherin-ZEB dual sensor in living cells, and ultimately to predict and alter cell fate during EMT in real time.

CONCLUSIONS

In this study, we use data assimilation to forecast cell fate during EMT. Using the data-assimilation approach incorporating multiplicative inflation and an optimal observation interval, EMT dynamics can be successfully reconstructed, and state transitions can be predicted before they occur, providing an opportunity to predict responses to biochemical perturbations in real time.

SUPPORTING MATERIAL

Supporting Material can be found online at https://doi.org/10.1016/j.bpj.2020.02.011.

AUTHOR CONTRIBUTIONS

M.J.M., M.J.H., E.M.C., C.A.L., and S.H.W. designed the research. M.J.M. carried out all simulations and analyzed the data. M.J.M., M.J.H., E.M.C., C.A.L., and S.H.W. wrote the article.

ACKNOWLEDGMENTS

This work was supported through funding from the National Institutes of Health/National Institute of General Medical Sciences R01GM122855 (S.H.W. and C.A.L.) and the National Science Foundation DCSD-1762803 (E.M.C. and M.J.H.).

REFERENCES

1. Thiery, J. P., H. Acloque, …, M. A. Nieto. 2009. Epithelial-mesenchymal transitions in development and disease. *Cell.* 139:871–890.
2. Brabletz, T., R. Kalluri, …, R. A. Weinberg. 2018. EMT in cancer. *Nat. Rev. Cancer.* 18:128–134.
3. Kalluri, R., and E. G. Neilson. 2003. Epithelial-mesenchymal transition and its implications for fibrosis. *J. Clin. Invest.* 112:1776–1784.
4. Miettinen, P. J., R. Ebner, …, R. Derynck. 1994. TGF-beta induced transdifferentiation of mammary epithelial cells to mesenchymal cells: involvement of type I receptors. *J. Cell Biol.* 127:2021–2036.
5. Xu, J., S. Lamouille, and R. Derynck. 2009. TGF-β-induced epithelial to mesenchymal transition. *Cell Res.* 19:156–172.

6. Griggs, L. A., N. T. Hassan, …, C. A. Lemmon. 2017. Fibronectin fibrils regulate TGF-β1-induced epithelial-mesenchymal transition. *Matrix Biol.* 60–61:157–175.

7. Scott, L. E., S. H. Weinberg, and C. A. Lemmon. 2019. Mechanochemical signaling of the extracellular matrix in epithelial-mesenchymal transition. *Front. Cell Dev. Biol.* 7:135.

8. Zhang, J., X.-J. Tian, …, J. Xing. 2014. TGF-β-induced epithelial-to-mesenchymal transition proceeds through stepwise activation of multiple feedback loops. *Sci. Signal.* 7:ra91.

9. Sha, Y., D. Haensel, …, Q. Nie. 2019. Intermediate cell states in epithelial-to-mesenchymal transition. *Phys. Biol.* 16:021001.

10. Pastushenko, I., and C. Blanpain. 2019. EMT transition states during tumor progression and metastasis. *Trends Cell Biol.* 29:212–226, Published online December 26, 2018.

11. Jolly, M. K., C. Ward, …, S. S. Sohal. 2018. Epithelial-mesenchymal transition, a spectrum of states: role in lung development, homeostasis, and disease. *Dev. Dyn.* 247:346–358.

12. Hong, T., K. Watanabe, …, X. Dai. 2015. An Ovol2-Zeb1 mutual inhibitory circuit governs bidirectional and multi-step transition between epithelial and mesenchymal states. *PLoS Comput. Biol.* 11:e1004569.

13. Nieto, M. A., R. Y.-J. Huang, …, J. P. Thiery. 2016. EMT: 2016. *Cell.* 166:21–45.

14. Lambert, A. W., D. R. Pattabiraman, and R. A. Weinberg. 2017. Emerging biological principles of metastasis. *Cell.* 168:670–691.

15. Tam, W. L., and R. A. Weinberg. 2013. The epigenetics of epithelial-mesenchymal plasticity in cancer. *Nat. Med.* 19:1438–1449.

16. Ruscetti, M., E. L. Dadashian, …, H. Wu. 2016. HDAC inhibition impedes epithelial-mesenchymal plasticity and suppresses metastatic, castration-resistant prostate cancer. *Oncogene.* 35:3781–3795.

17. Bhatia, S., J. Monkman, …, E. W. Thompson. 2019. Interrogation of phenotypic plasticity between epithelial and mesenchymal states in breast cancer. *J. Clin. Med.* 8:E893.

18. Tripathi, S., P. Chakraborty, …, M. K. Jolly. 2020. A mechanism for epithelial-mesenchymal heterogeneity in a population of cancer cells. *PLoS Comput. Biol.* 16:e1007619.

19. Tian, X.-J., H. Zhang, and J. Xing. 2013. Coupled reversible and irreversible bistable switches underlying TGFβ-induced epithelial to mesenchymal transition. *Biophys. J.* 105:1079–1089.

20. Sobie, E. A. 2009. Parameter sensitivity analysis in electrophysiological models using multivariable regression. *Biophys. J.* 96:1264–1274.

21. Sarkar, A. X., and E. A. Sobie. 2010. Regression analysis for constraining free parameters in electrophysiological models of cardiac cells. *PLoS Comput. Biol.* 6:e1000914.

22. Sarkar, A. X., D. J. Christini, and E. A. Sobie. 2012. Exploiting mathematical models to illuminate electrophysiological variability between individuals. *J. Physiol.* 590:2555–2567.

23. Deshiere, A., E. Duchemin-Pelletier, …, O. Filhol. 2013. Unbalanced expression of CK2 kinase subunits is sufficient to drive epithelial-to-mesenchymal transition by Snail1 induction. *Oncogene.* 32:1373–1383.

24. Taube, J. H., J. I. Herschkowitz, …, S. A. Mani. 2010. Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes. *Proc. Natl. Acad. Sci. USA.* 107:15449–15454.

25. Hesling, C., L. Fattet, …, R. Rimokh. 2011. Antagonistic regulation of EMT by TIF1γ and Smad4 in mammary epithelial cells. *EMBO Rep.* 12:665–672.

26. Zhang, J., H. Chen, …, J. Xing. 2019. Spatial clustering and common regulatory elements correlate with coordinated gene expression. *PLoS Comput. Biol.* 15:e1006786.

27. Toneff, M. J., A. Sreekumar, …, J. M. Rosen. 2016. The Z-cad dual fluorescent sensor detects dynamic changes between the epithelial and mesenchymal cellular states. *BMC Biol.* 14:47.

28. Hunt, B. R., E. J. Kostelich, and I. Szunyogh. 2007. Efficient data assimilation for spatiotemporal chaos: a local ensemble transform Kalman filter. *Physica D.* 230:112–126.

29. Ghil, M., and P. Malanotte-Rizzoli. 1991. Data assimilation in meteorology and oceanography. *Adv. Geophys.* 33:141–266.

30. Houtekamer, P., and F. Zhang. 2016. Review of the ensemble Kalman filter for atmospheric data assimilation. *Mon. Weather Rev.* 144:4489–4532.

31. Evensen, G. 2009. The ensemble Kalman filter for combined state and parameter estimation. *IEEE Control Systems.* 29:83–104.

32. Kostelich, E. J., Y. Kuang, …, M. C. Preul. 2011. Accurate state estimation from uncertain data and models: an application of data assimilation to mathematical models of human brain tumors. *Biol. Direct.* 6:64.

33. Hamilton, F., T. Berry, …, T. Sauer. 2013. Real-time tracking of neuronal network structure using data assimilation. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 88:052715.

34. Hamilton, F., J. Cressman, …, T. Sauer. 2014. Reconstructing neural dynamics using data assimilation with multiple models. *EPL.* 107:68005.

35. Hamilton, F., T. Berry, and T. Sauer. 2018. Tracking intracellular dynamics through extracellular measurements. *PLoS One.* 13:e0205031.

36. Ullah, G., and S. J. Schiff. 2010. Assimilating seizure dynamics. *PLoS Comput. Biol.* 6:e1000776.

37. Ullah, G., and S. J. Schiff. 2009. Tracking and control of neuronal Hodgkin-Huxley dynamics. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 79:040901.

38. Hoffman, M. J., N. S. LaVigne, …, E. M. Cherry. 2016. Reconstructing three-dimensional reentrant cardiac electrical wave dynamics using data assimilation. *Chaos.* 26:013107.

39. LaVigne, N. S., N. Holt, …, E. M. Cherry. 2017. Effects of model error on cardiac electrical wave state reconstruction using data assimilation. *Chaos.* 27:093911.

40. Moye, M. J., and C. O. Diekman. 2018. Data assimilation methods for neuronal state and parameter estimation. *J. Math. Neurosci.* 8:11.

41. Sedigh-Sarvestani, M., D. J. Albers, and B. J. Gluckman. 2012. Data assimilation of glucose dynamics for use in the intensive care unit. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2012:5437–5440.

42. Sedigh-Sarvestani, M., S. J. Schiff, and B. J. Gluckman. 2012. Reconstructing mammalian sleep dynamics with data assimilation. *PLoS Comput. Biol.* 8:e1002788.

43. Kalnay, E., H. Li, …, J. Ballabrera-Poy. 2007. 4-D-Var or ensemble Kalman filter? *Tellus, Ser. A, Dyn. Meterol. Oceanogr.* 59:758–773.

44. Szunyogh, I., E. J. Kostelich, …, J. A. Yorke. 2008. A local ensemble transform Kalman filter data assimilation system for the NCEP global model. *Tellus, Ser. A, Dyn. Meterol. Oceanogr.* 60:113–130.

45. Miyoshi, T., and M. Kunii. 2011. The local ensemble transform Kalman filter with the weather research and forecasting model: experiments with real observations. *Pure Appl. Geophys.* 169:321–333.

46. Hoffman, M. J., S. J. Greybush, …, I. Szunyogh. 2010. An ensemble Kalman filter data assimilation system for the Martian atmosphere: implementation and simulation experiments. *Icarus.* 209:470–481.

47. Hoffman, M. J., T. Miyoshi, …, R. Murtugudde. 2012. An advanced data assimilation system for the Chesapeake bay: performance evaluation. *J. Atmos. Ocean. Technol.* 29:1542–1557.

48. Kadakia, N., E. Armstrong, …, H. D. Abarbanel. 2016. Nonlinear statistical data assimilation for HVC_{RA} neurons in the avian song system. *Biol. Cybern.* 110:417–434.

49. Lu, M., M. K. Jolly, …, E. Ben-Jacob. 2013. MicroRNA-based regulation of epithelial-hybrid-mesenchymal fate determination. *Proc. Natl. Acad. Sci. USA.* 110:18144–18149.

50. Munoz, L. M., and N. F. Otani. 2013. Kalman filter based estimation of ionic concentrations and gating variables in a cardiac myocyte model. *In* Computing in Cardiology 2013. A. Murray, ed. IEEE, pp. 53–56.

51. Meliza, C. D., M. Kostuk, …, H. D. Abarbanel. 2014. Estimating parameters and predicting membrane voltages with conductance-based neuron models. *Biol. Cybern.* 108:495–516.

52. Wang, J., D. Breen, …, G. Cauwenberghs. 2016. Data assimilation of membrane dynamics and channel kinetics with a neuromorphic integrated circuit. *In* 2016 IEEE Biomedical Circuits and Systems Conference (BioCAS). IEEE, pp. 584–587.

53. Basu, S., S. Cheriyamundath, and A. Ben-Ze'ev. 2018. Cell-cell adhesion: linking Wnt/β-catenin signaling with partial EMT and stemness traits in tumorigenesis. *F1000 Res.* 7:F1000 Faculty Rev-1488.

54. Hua, K., Y. Li, …, H. Jin. 2018. Haemophilus parasuis infection disrupts adherens junctions and initializes EMT dependent on canonical Wnt/β-catenin signaling pathway. *Front. Cell. Infect. Microbiol.* 8:324.

55. Scott, L. E., L. A. Griggs, …, S. H. Weinberg. 2019. A predictive model of intercellular tension and cell-matrix mechanical interactions in a multicellular geometry. *bioRxiv* https://doi.org/10.1101/701037.

56. Bocci, F., L. Gearhart-Serna, …, M. K. Jolly. 2019. Toward understanding cancer stem cell heterogeneity in the tumor microenvironment. *Proc. Natl. Acad. Sci. USA.* 116:148–157.

57. Salgia, R., I. Mambetsariev, …, M. Sattler. 2018. Modeling small cell lung cancer (SCLC) biology through deterministic and stochastic mathematical models. *Oncotarget.* 9:26226–26242.

58. Metzcar, J., Y. Wang, …, P. Macklin. 2019. A review of cell-based computational modeling in cancer biology. *JCO Clin. Cancer Inform.* 3:1–13.

59. Law, K., A. Stuart, and K. Zygalakis. 2015. Data Assimilation. Springer, Cham, Switzerland.

60. Asch, M., M. Bocquet, and M. Nodet. 2016. Data Assimilation: Methods, Algorithms, and Applications. SIAM, Philadelphia, PA.

61. Barone, A., F. Fenton, and A. Veneziani. 2017. Numerical sensitivity analysis of a variational data assimilation procedure for cardiac conductivities. *Chaos.* 27:093930.