A Population-Informed Particle Filter for Robust Physiological Monitoring using Low-Information Time-Series Measurements

Ali Tivay and Jin-Oh Hahn, *Senior Member, IEEE*

***Abstract*— *Objective:* To present the population-informed particle filter (PIPF), a novel filtering approach that incorporates past experiences with patients into the filtering process to provide reliable beliefs about a new patient's physiological state. *Methods:* To derive the PIPF, we formulate the filtering problem as recursive inference on a probabilistic graphical model, which includes representations for the pertinent physiological dynamics and the hierarchical relationship between past and present patient characteristics. Then, we provide an algorithmic solution to the filtering problem using Sequential Monte-Carlo techniques. To demonstrate the merits of the PIPF approach, we apply it to a case study of physiological monitoring for hemodynamic management. *Results:* The PIPF approach could provide reliable beliefs about the likely values and uncertainties associated with a patient's unmeasured physiological variables (e.g., hematocrit and cardiac output), characteristics (e.g., tendency for atypical behavior), and events (e.g., hemorrhage) given low-information measurements. *Conclusion:* The PIPF shows promise in the presented case study, and may have applications to a wider range of real-time monitoring problems with limited measurements. *Significance:* Forming reliable beliefs about a patient's physiological state is an essential aspect of algorithmic decision-making in medical care settings. Hence, the PIPF may serve as a solid basis for designing interpretable and context-aware physiological monitoring, medical decision-support, and closed-loop control algorithms.**

***Index Terms*—Particle Filter, Generative Model, Recursive Inference, Physiological Monitoring, Critical Care, Hemodynamic Management**

I. Introduction

PHYSIOLOGICAL monitoring systems are foundational tools for patient care that provide continuous information about a patient's vital physiological variables, helping practitioners in deriving insight into the patient's health condition and making informed therapeutic decisions [1]. The need for physiological monitoring also spans the non-clinical domain, where a wide range of monitoring products (e.g., wearables and consumer electronics) aim to provide users with insight into their health and bodily performance [2]. In addition, in recent years, there has been considerable research interest in the area of autonomous medical care systems [3]–[6], where physiological decision-support and closed-loop control algorithms are built to assist users in making therapeutic or lifestyle decisions based on monitoring results. Such a trend makes it even more necessary to develop high-fidelity and reliable physiological monitoring systems.

As a fundamental challenge in physiological monitoring, the physiological variables that are relevant to decision-making are not always directly measurable in practice. As a result, to be useful, monitoring systems must have mechanisms to continuously infer unmeasured physiological variables from measured ones. In the context of engineering systems, *filtering algorithms* are excellent candidates for such purposes [7], [8]. These algorithms typically leverage an underlying model of the studied system (which could be mechanistic or black box) to recursively process measurement signals and continuously infer unmeasured variables. The Kalman Filter (KF) is one of the most well-known and widely used filtering algorithms in the engineering domain, which can utilize a linear model of the system for estimation purposes [9], [10]. The Extended Kalman Filter (EKF) and the Unscented Kalman Filter (UKF) are two extensions to the KF algorithm that use model approximation techniques to allow for a nonlinear model of the system to be used for estimation [11]–[13]. More generally, Bayesian filtering provides a concrete framework for reasoning about filtering problems, where beliefs about the possible values of a system's unmeasured variables are expressed in the form of probability distributions [14]. In this framework, principled mathematical procedures exist to refine existing (i.e., prior) beliefs according to incoming data in order to turn them into updated (i.e., posterior) beliefs about unmeasured variables [15], [16]. However, these mathematical operations are often not analytically tractable, which fostered the development of a wide range of approximate approaches that provide solutions to such problems in practical applications. These approaches can be divided into two broad categories in terms of their approximation technique. Sequential Monte-Carlo (SMC) filtering approaches approximate belief distributions using (large) collections of weighted samples (i.e., particles), allowing for stochastic and/or nonlinear models to be used in

This work was supported by the U.S. National Science Foundation CAREER Award (Grant CNS-1748762).

A.Tivay and J.O. Hahn are with the Department of Mechanical Engineering, University of Maryland, College Park, MD 20742, USA (correspondence e-mail: jhahn12@umd.edu).

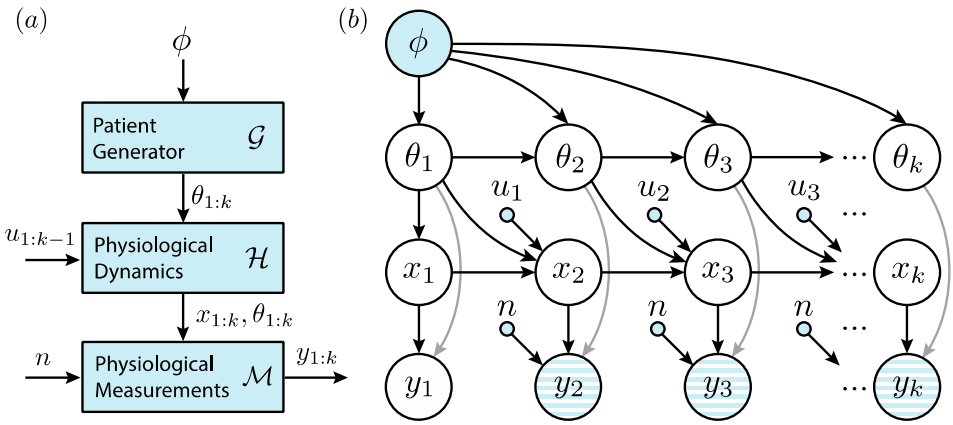

Fig. 1. Schematic representations for (a) the hierarchical structure of the generative physiological model; and (b) the sequential relationship between the variables of interest in the population-informed filtering problem.

the filtering process with relatively high fidelity [17]–[21], while Variational Filtering (VF) approaches approximate belief distributions using tunable distribution models, allowing for the filtering problem to be converted into an optimization problem [22]–[25].

Despite the demonstrated success of filtering algorithms in estimating unmeasured variables in many engineering systems, unique filtering challenges still remain in the physiological monitoring domain, namely: the available measurements typically contain limited information about the unmeasured variables. For instance, in the context of hemodynamic management for critically ill patients, beliefs about the patient's blood volume/composition and cardiovascular function are expected to facilitate a practitioner's (or an algorithm's) decisions in administering therapy (e.g., fluid or drug infusions). However, clinically available measurements are limited to blood pressure and (in rare cases) intermittent cardiac output and hematocrit correlates, which are also typically affected by high levels of noise and artifacts. Moreover, a patient's physiological dynamics are typically not excited by the therapy in clinical settings, further limiting the information content of the available measurements. Such intermittent and low-information measurements limit the applicability of many established filtering algorithms (and underlying physiological models) to the task of characterizing patients and providing reliable estimates of their physiological state.

To address this challenge, in this work, we propose the population-informed particle filter (PIPF), a novel filtering approach that leverages a *generative physiological model* [26] to incorporate past experiences with patients into the filtering process in order to provide reliable beliefs about a new patient's physiological state. To derive the PIPF, we formulate the filtering problem as recursive Bayesian inference on a probabilistic graphical model, which includes representations for the pertinent physiological dynamics and the hierarchical relationship between past and present patient characteristics. Then, we provide an algorithmic solution to the filtering problem based on SMC (i.e., particle-based) techniques. To demonstrate the potential merits and limitations of the PIPF approach, we apply it to a case study of physiological monitoring for hemodynamic management.

This paper is organized as follows. Section II presents the algorithmic details of the PIPF approach. Section III presents the application of the PIPF to physiological monitoring in the context of hemodynamic management. Section IV presents and discusses the results. Section V concludes the paper with potential future directions.

II. Population-Informed Particle Filtering

In this section, we present the population-informed particle filter (PIPF). First, we provide a review of relevant concepts from generative physiological modeling. Based on these concepts, we present the population-informed filtering scheme, where a generative physiological model informs a recursive Bayesian filter. Then, we provide an algorithmic solution to this problem using SMC techniques and show how this algorithm and methodology may be leveraged to create a robust model-based physiological monitoring system. Further details follow.

A. Generative Physiological Modeling

Physiological models can serve as a concrete source of physiological knowledge in the design and development of patient monitoring algorithms. Generative modeling is a promising approach to physiological modeling, where the inherent variability and stochasticity of a physiological system are fully embraced in model components that behave in stochastic but patterned ways. The objective of the generative physiological model is therefore to reproduce and predict the patterned randomness that is often observed in physiological datasets. As the proposed filtering approach relies heavily on an underlying generative model, in this section, we present a general family of generative models for physiological systems and provide an overview of procedures that can be used to characterize these models from data. We refer the readers to our prior work [26] for complete details on this topic.

The generative model considered in this work consists of a hierarchy of stochastic components that reflect the physiological data observed in a population of patients (see Fig. 1(a)). At the highest level in the hierarchy, a patient generator

model is tasked with generating variations in patient characteristics, which can be formalized as:

$$\theta_1 \sim \mathcal{G}_1(\phi) \quad (1)$$

$$\theta_k \sim \mathcal{G}_k(\theta_{k-1}, \phi) \quad (2)$$

where $\mathcal{G}_1$ is a component that instantiates virtual patients by generating patient characteristics, and $\mathcal{G}_k$ is a component that produces variations in a virtual patient's characteristics as time progresses. In this formulation, ϕ is the vector of parameters for the patient generator model, θ_k is the vector of patient characteristics at time k, and the symbol $\sim$ denotes sampling. At the second level, a physiological dynamics model is tasked with generating evolutions in the states of each virtual patient, which can be formalized as:

$$x_1 \sim \mathcal{H}_1(\theta_1) \quad (3)$$

$$x_k \sim \mathcal{H}_k(x_{k-1}, \theta_{k-1}, u_{k-1}) \quad (4)$$

where $\mathcal{H}$ is the physiological dynamics model, x_k represents the states of the virtual patient at time k, and u_k represents the known inputs/therapies given to the virtual patient at time k. Finally, at the third level, a physiological measurement model generates observations from the state and/or the characteristics of the patient, which can be formalized as:

$$y_k \sim \mathcal{M}(x_k, \theta_k, n) \quad (5)$$

where $\mathcal{M}$ is the physiological measurement model, n is a vector of parameters for this model, and y_k is the vector of virtual observations generated at time k.

Given a dataset containing physiological data from a cohort of patients, we are interested in inferring the unknown parameters of the generative model in (1)-(5) such that the model captures the characteristics of the dataset. We recently showed in [26] that an effective solution to this (often intractable) problem can be obtained using variational Bayesian inference methods [27], where the most-likely values and the uncertainties associated with the parameters of the generative model (i.e., ϕ, n) are computed through stochastic optimization. These inferred parameters can in turn be used with the generative model to generate virtual datasets with similar distribution to real data. In other words, the resulting generative model is equipped with the knowledge needed to instantiate virtual patients, generate paths for patient characteristics, produce state evolutions in response to given stimuli, and generate realistic physiological measurements. In this work, we are interested in utilizing this encoded knowledge to inform a filtering algorithm's real-time perception of a patient, especially when only intermittent and low-information measurements are available from the patient. Therefore, in the next section, we formulate a filtering problem where the filter is informed by a generative model.

B. Population-Informed Filtering

Filtering (as referred to in this work) is the process of forming and updating beliefs about the unmeasured variables of a dynamic system using its measured variables. In the context of physiological monitoring, filtering may be used to form and update beliefs about the unmeasured states and characteristics of a patient using physiological measurements that arrive over time. Filtering processes typically operate based on an underlying model of the studied system. In this section, we formulate the problem of *population-informed filtering*, which is a filtering process that is informed by a generative physiological model.

To formulate the population-informed filtering problem, we adopt a Bayesian view of filtering, where beliefs are represented by probability density functions. In this setting, the objective of filtering is to obtain the following *posterior probability density function* at each time k:

$$p(x_{1:k}, \theta_{1:k} | y_{1:k}, \phi, n) = \frac{p(x_{1:k}, \theta_{1:k}, y_{1:k} | \phi, n)}{p(y_{1:k} | \phi, n)} \quad (6)$$

This density represents our beliefs about the unmeasured states and characteristics of a patient up to time k (denoted by $x_{1:k}$ and $\theta_{1:k}$ respectively), given the patient's physiological measurements up to time k (denoted by $y_{1:k}$) and the information encoded in the parameters of the generative model ϕ, n. As shown on the right-hand-side of (6), it is conceptually possible to obtain the posterior density from the *joint density* shown in the numerator and the *marginal density* shown in the denominator. The joint density represents the relationship between the measured and unmeasured variables in the filtering problem, while the marginal density is obtained by integrating the joint density over the unmeasured variables.

Using the generative model in (1)-(5) to inform the filtering process imposes a specific structure on the joint density. This structure is shown in Fig. 1(b) as a graphical representation. In this representation, the joint density has a hierarchical structure, where the patient generator parameters ϕ affect the patient's characteristics $\theta_{1:k}$, while the patient's characteristics affect the patient's state evolutions $x_{1:k}$. In turn, the patient's characteristics, state evolutions, and measurement parameters n affect the generated measurements $y_{1:k}$. The graphical representation also shows that the joint density consists of a sequential structure, where variables at time k are affected by variables at time $k-1$. This implies a recursive relationship between joint densities at times k and $k-1$, which can be formalized as:

$$p(x_{1:k}, \theta_{1:k}, y_{1:k} | \phi, n) = p(x_{1:k-1}, \theta_{1:k-1}, y_{1:k-1} | \phi, n) \mathcal{R}_{k-1:k} \quad (7)$$

$$\mathcal{R}_{k-1:k} = p(y_k | x_k, \theta_k, n) p(x_k | x_{k-1}, \theta_{k-1}, u_{k-1}) p(\theta_k | \theta_{k-1}, \phi) \quad (8)$$

where $\mathcal{R}_{k-1:k}$ is the ratio between joint densities at times k and $k-1$. The first multiplicative term in the ratio is the density associated with generating the measurement y_k given the current states x_k and characteristics θ_k; the second term is the density associated with generating a transition to the states x_k given the previous states x_{k-1}, characteristics θ_{k-1}, and inputs/therapies u_{k-1}; and the third term is the density associated with generating a transition to the characteristics θ_k given the previous characteristics θ_{k-1}.

The relationship described in (7)-(8) results in an important recursive relationship between posterior densities at times k and $k-1$. This relationship can be written as:

$$p(x_{1:k}, \theta_{1:k} | y_{1:k}, \phi, n) = p(x_{1:k-1}, \theta_{1:k-1} | y_{1:k-1}, \phi, n) \frac{\mathcal{R}_{k-1:k}}{p(y_k | y_{1:k-1}, \phi, n)} \quad (9)$$

which suggests that it is conceptually possible to multiply the posterior density at time $k-1$ by an update term (i.e., the

fraction on the right-hand side) to obtain the posterior density at time k. In other words, performing this update recursively would create a recursive filtering process, where previous beliefs about a patient's unmeasured states and characteristics can be updated according to the behavior of the generative model and any newly available measurements y_k. However, depending on the underlying generative model, this update procedure may often prove analytically intractable. In the next section, we adopt a Sequential Monte-Carlo (i.e., particle-based) approach to derive a practical solution to this filtering problem.

C. A Sequential Monte-Carlo Solution

To derive a practical solution to the population-informed filtering problem presented in (9), we adopt a Sequential Monte-Carlo approach [20], [21], which is a sampling-based approach especially suited to solving inference problems with substantial nonlinearities, complex/multi-modal beliefs, and high levels of uncertainty, most of which are likely to arise in physiological monitoring applications. To perform this derivation, we first consider a *proposal density* of the following form:

$$q(x_{1:k}, \theta_{1:k}) = q(x_{1:k-1}, \theta_{1:k-1}) q(x_k, \theta_k | x_{k-1}, \theta_{k-1}) \quad (10)$$

which is designed to have a sequential structure, be easy to sample from, and have non-zero density wherever the posterior density (9) is expected to have non-zero density. The function of this proposal density is to propose beliefs about the unmeasured variables of the filtering problem. Multiplying and dividing (9) by (10) and rearranging the result yields the following variant of the recursive relationship between the posterior densities:

$$p(x_{1:k}, \theta_{1:k} | y_{1:k}, \phi, n) = p(x_{1:k-1}, \theta_{1:k-1} | y_{1:k-1}, \phi, n) \frac{\alpha(x_{k-1:k}, \theta_{k-1:k}) q(x_k, \theta_k | x_{k-1}, \theta_{k-1})}{p(y_k | y_{1:k-1}, \phi, n)} \quad (11)$$

where $\alpha(x_{k-1:k}, \theta_{k-1:k})$ is a *weighting function* defined as:

$$\alpha(x_{k-1:k}, \theta_{k-1:k}) = \frac{\mathcal{R}_{k-1:k}}{q(x_k, \theta_k | x_{k-1}, \theta_{k-1})} \quad (12)$$

To perform the recursive filtering procedure described by (11)-(12), we start from the process of drawing samples from the proposal density in (10):

$$[x, \theta]_k^i \sim q\left(x_k, \theta_k | [x, \theta]_{k-1}^i\right) \quad (13)$$

where $[x, \theta]_k^i$ denotes a sample (at time k and indexed by i) from the proposal density, and it follows from the sequential structure of the proposal density that samples at time k can be generated recursively using samples at time $k-1$. Given N such samples, the proposal density itself can be expressed using the following equation:

$$\hat{q}\left(x_k, \theta_k | [x, \theta]_{k-1}^i\right) = \frac{1}{N} \sum_{i=1}^{N} \delta[x, \theta]_k^i \quad (14)$$

where $\delta[x, \theta]_k^i$ denotes the Dirac's delta function positioned at the sample $[x, \theta]_k^i$, and the equation indicates that the proposal density can be approximately represented by a collection of N samples $[x, \theta]_k^i$ drawn from it. Given the samples, the approximation in (14) can be substituted into (11) over the time span $1{:}\,k$ to yield the following sample-based expression for the posterior density:

$$\hat{p}(x_{1:k}, \theta_{1:k} | y_{1:k}, \phi, n) = \sum_{i=1}^{N} w_k^i \delta[x, \theta]_{1:k}^i \quad (15)$$

Algorithm 1. Population-Informed Particle Filtering (PIPF)

$k \leftarrow 1$ {initialize time}

$\log w_1^i \leftarrow \log(1/N)$

{initialize N weights, indexed by i}

$\theta_1^i \sim \mathcal{G}_1(\phi),\ x_1^i \sim \mathcal{H}_1(\theta_1^i)$

{generate N samples, indexed by i; see (1), (3)}

Repeat:

$k \leftarrow k + 1$ {increment time}

$\theta_k^i \sim \mathcal{G}_k(\theta_{k-1}^i, \phi)$

{generate transition to characteristic samples; see (2)}

$x_k^i \sim \mathcal{H}_k(x_{k-1}^i, \theta_{k-1}^i, u_{k-1})$

{generate transition to state samples; see (4)}

$\log \alpha^i \leftarrow \log p(y_k | x_k^i, \theta_k^i, n)$

{get log-likelihood of samples against data; see (12)}

$\log \overline{w}_k^i \leftarrow \log w_{k-1}^i + \log \alpha^i$

{update weights, un-normalized; see (16)}

$\log \overline{w}_k^* \leftarrow \max_i(\log \overline{w}_k^i)$

{get maximum weight}

$\log w_k^i \leftarrow \log \overline{w}_k^i - \log \overline{w}_k^* - \log \sum_i \exp(\log \overline{w}_k^i - \log \overline{w}_k^*)$

{normalize weights}

$[x, \theta]_k^i \leftarrow \mathcal{Z}(\log w_k^i, [x, \theta]_k^i)$

{resample to reset weights; see [21]}

$\log w_k^i \leftarrow \log\left(\frac{1}{N}\right)$

{reset weights}

Output: $\log w_k^i$, $[x, \theta]_k^i$

{provide up-to-date beliefs for time k}

where the posterior density is expressed in terms of weighted samples from the proposal density, and w_k^i denotes the weight for sample i at time k. Each sample path $[x, \theta]_{1:k}^i$ describes a proposed path for the patient's states and characteristics, and the weight w_k^i represents how probable the path is (relative to other sample paths) according to the generative model and the patient's physiological measurements up to time k. Each weight at time k can be computed recursively from its counterpart at time $k-1$ using the following relationship:

$$w_k^i = \frac{w_{k-1}^i \alpha([x, \theta]_{k-1:k}^i)}{\sum_{i=1}^{N} w_{k-1}^i \alpha([x, \theta]_{k-1:k}^i)} \quad (16)$$

where w_{k-1}^i denotes the weight for sample i at time $k-1$, and $\alpha([x, \theta]_{k-1:k}^i)$ is the weighting function described in (12), evaluated at sample i for times k and $k-1$. Overall, the relationship described in (16) allows for the efficient calculation of the posterior density using recursive steps, namely: drawing N samples at time k from the proposal density (13) using the already available N samples from time $k-1$; evaluating the weighting function (12) using each of the N samples at times k and $k-1$; and calculating the new weights

using (16), thereby obtaining the most up-to-date beliefs at time k. Realizing this procedure in practice, however, necessitates a few extra steps, which are presented in more detail in the next section.

D. The Population-Informed Particle Filtering (PIPF) Algorithm

As presented in Section II.C, beliefs about a patient's unmeasured states and characteristics can be expressed and recursively updated using a collection of weighted samples (i.e., particles). In this section, we leverage this principle in conjunction with additional weight normalization and resampling techniques to build a practical procedure for PIPF. Algorithm 1 shows an overview of this procedure. In this algorithm, we use the generative physiological model in (1)-(5) to generate proposal samples. As a result, the weighting function used in (16) reduces to $\alpha\left([x,\theta]_{k-1:k}^{i}\right) \triangleq p\left(y_k \middle| x_k^i, \theta_k^i, n\right)$. The filtering procedure begins by using the generative model to generate (many) proposed samples for the patient's baseline states and characteristics. Then, at each time increment, the generative model proposes transitions to these states and characteristics, and the resulting samples are evaluated against data at the time to update the sample weights. The weight updates are performed in logarithmic scale and the weight normalizations are performed using the log-sum-exp trick [28] to achieve higher numerical accuracy and stability. In addition, weight variability is measured at each time increment, and if necessary, the samples are resampled using a systematic resampling [21] technique (denoted by $\mathcal{Z}$ in Algorithm 1). This technique redraws each sample with a probability proportional to its weight, thereby allowing for a resetting of the weights. Overall, at each time k, this algorithm provides up-to-date beliefs about a patient's unmeasured variables in the form of a collection of weighted samples (i.e., particles).

III. Application to Physiological Monitoring for Hemodynamic Management

Hemodynamic management is an essential aspect of care for critically ill patients, where the performance of a patient's cardiovascular system is monitored and, if necessary, therapies are administered to ensure adequate blood circulation [29], [30]. To demonstrate the merits and limitations of PIPF, we use this approach to build a monitoring algorithm for a typical critical care scenario where hypovolemia (i.e., low circulating blood volume) is treated with fluid infusions [31]. In this scenario, the monitoring algorithm receives a stream of blood pressure and fluid infusion data and processes this information to form real-time beliefs about a patient's unmeasured physiological variables (e.g., hematocrit and cardiac output), characteristics (e.g., the tendency for atypical behavior), and events (e.g., hemorrhage). These real-time beliefs can in turn inform a practitioner's decision to administer fluids, and furthermore serve as a foundation for building automated decision-support and closed-loop control algorithms for hypovolemia treatment. This section presents the details of applying the PIPF approach to this problem and describes the methods and datasets used to evaluate the results. Further details follow.

A. Generative Modeling for Hemodynamic Management

As presented in Section II, the PIPF approach relies on a generative physiological model to form and update beliefs about the unmeasured states and characteristics of a patient. In this section, we describe a generative model of the physiological phenomena relevant to the treatment of hypovolemia with fluid infusions. This model will be used in subsequent sections to build a monitoring algorithm for hypovolemia treatment.

1) Patient Generator Model: As presented in Section II.A, the generative physiological model consists of a hierarchy of stochastic components that reflect the physiological behaviors observed in a population of patients. At the highest level in the hierarchy, a patient generator model $\mathcal{G}$ generates virtual patients with a variety of characteristics. In the context of hypovolemia treatment, we consider a patient generator model of the following form:

$$\theta_1 = \phi_\mu + \phi_L \epsilon, \epsilon \sim \mathcal{N}(0, I) \tag{17}$$

$$\theta_k = (1-\epsilon_b)\theta_{k-1} + \epsilon_b(\phi_\mu + \phi_L \epsilon),\ \epsilon_b \sim \mathcal{B}(1, \gamma_\theta),\ \epsilon \sim \mathcal{N}(0, I) \tag{18}$$

In this model, first, the characteristics θ_1 of each virtual patient are instantiated in (17) by drawing samples from a full-covariance Gaussian distribution, where ϕ_μ determines the center of the distribution, ϕ_L is the Cholesky decomposition of the distribution's covariance matrix (i.e., $\phi_\Sigma = \phi_L \phi_L^T$), and ϵ is a sample from the standard Gaussian distribution of appropriate dimension. Then, over time, each patient's characteristics θ_k follow the relationship in (18), where ϵ_b is a sample from the Bernoulli distribution (denoted by $\mathcal{B}$), which produces $\epsilon_b = 1$ with a probability of γ_θ, and $\epsilon_b = 0$ with a probability of $1 - \gamma_\theta$. The probability γ_θ (which is chosen to be small) acts as a forgetting factor for the characteristics of each instantiated virtual patient. In other words, each virtual patient will retain its characteristics over time, except for a small chance of transitioning to different characteristics drawn from the virtual patient generator in (17). For the purpose of hypovolemia treatment modeling, we define a vector of patient characteristics as follows:

$$\theta_k = \left[\, v_0 \;\; H_0 \;\; Q_0 \;\; P_{a0} \;\; K_v \;\; K_a/K_v \;\; K_p \;\; \alpha_I \;\; \alpha_H \;\; K_h \;\; \tau_R \;\; K_R \;\; \beta_v \;\; K_Q \,\right]_k \tag{19}$$

which contains $n_p = 14$ physiological parameters to be elaborated on later in this section. Overall, this patient generator model is built to inform a PIPF-based monitoring algorithm about the characteristics that are likely to occur in the patient population (along with their probability of occurrence) and furthermore allow for the algorithm to adapt to potential changes in a patient's characteristics over time.

2) Physiological Dynamics Model: At the second level in the hierarchical model in Section II.A, a physiological dynamics model $\mathcal{H}$ generates state evolutions for each virtual patient. For this purpose, we utilize a dynamic model of the physiological phenomena relevant to hypovolemia treatment from our previous work [26], [32]. In this section, we present an overview of this model in discretized form. The model consists of macro-state components that represent blood

circulation and the mechanisms that affect blood circulation in the context of hypovolemia treatment.

To obtain a macro-state model of blood circulation, we consider equations of the following form:

$$[v_a]_k = [v_a]_{k-1} + \delta t\,[Q - (P_a - P_v)/R - J_H - J_F]_{k-1} \quad (20)$$

$$[v_v]_k = [v_v]_{k-1} + \delta t\,[-Q + (P_a - P_v)/R + J_I]_{k-1} \quad (21)$$

$$[v_r]_k = [v_r]_{k-1} + \delta t\,[-J_H H]_{k-1} \quad (22)$$

where v_a and v_v denote arterial and venous blood volumes, v_r is the total red blood cell volume, δt is the time increment between k and $k-1$ instances, Q is the cardiac output, P_a and P_v denote mean arterial and venous blood pressures, R is the systemic vascular resistance, $H = v_r/(v_a + v_v)$ is the hematocrit, J_F is the net rate of fluid exchange with the interstitial space, J_I is the rate of fluid infusion, and J_H is the rate of blood loss. In these equations, mean arterial pressure is related to arterial volume through $P_a = P_{a0} + K_a(v_a - v_{a0})$, and mean venous pressure is related to venous volume through $P_v = P_{v0} + K_v(v_v - v_{v0})$, where P_{a0}, P_{v0} are baseline (i.e., un-perturbed) arterial and venous blood pressures, v_{a0}, v_{v0} are baseline arterial and venous blood volumes, and K_a, K_v denote arterial and venous elastances in the patient.

To obtain a macro-state model of the mechanisms that affect blood circulation in the context of hypovolemia treatment, we consider: (i) fluid exchange with the interstitial space, (ii) changes in systemic vascular resistance (e.g., through vasoconstriction or vasodilation), and (iii) changes in cardiac output (e.g., through changes to heart rate and contractility). Fluid exchange with the interstitial space is represented by:

$$J_F = K_p(v - v_0 - r_F) \quad (23)$$

$$[r_F]_k = [r_F]_{k-1} + \delta t\,[J_I/(1+\alpha_I) - J_H/(1+\alpha_H)]_{k-1} \quad (24)$$

where r_F is the steady-state change in blood volume after the exchange, $v = v_a + v_v$ is the total blood volume, $v_0 = v_{a0} + v_{v0}$ is the baseline total blood volume, K_p modulates the speed of exchange, and α_I, α_H determine the fraction of fluid infusion and blood loss that are compensated in the exchange. Changes in systemic vascular resistance are represented by:

$$R = R_0 + K_h(H - H_0) + s_R \quad (25)$$

$$[s_R]_k = [s_R]_{k-1} + \delta t\,[-s_R/\tau_R - K_R(P_a - P_{a0})/\tau_R]_{k-1} \quad (26)$$

where s_R is the change in resistance prompted by the body's compensatory mechanisms, K_R and τ_R modulate the characteristics of compensation, $R_0 = (P_{a0} - P_{v0})/Q_0$ is the baseline resistance, Q_0 is the baseline cardiac output, H_0 is the baseline hematocrit, and K_h modulates the effect of blood dilution on resistance. Changes in cardiac output are represented by:

$$Q = Q_0 + \beta_v(P_v - P_{v0}) + s_Q \quad (27)$$

$$[s_Q]_k = [s_Q]_{k-1} + \delta t\,[-K_Q(Q - Q_0)]_{k-1} \quad (28)$$

where s_Q is the change in cardiac output prompted by the body's compensatory mechanisms, K_Q modulates the characteristics of compensation, and β_v modulates the effect of variations in mean venous pressure on cardiac output. The physiological model described in (20)-(28) is intended to inform a PIPF-based monitoring algorithm about the physiological dynamics and behaviors that could occur in a patient, thereby giving the algorithm a basis to form and adjust its beliefs about the patient's state over time.

3) Hemorrhage Events Model: The model described in Section III.A.2 is built to represent the physiological response of a patient to fluid infusions (i.e., the rate J_I) and hemorrhage (i.e., the rate J_H). In the context of hypovolemia treatment, fluid infusions are typically known, as they would be administered to the patient by a caregiver (or an algorithm). In contrast, the presence of hemorrhage may be unknown in many cases. In such cases, a model should be devised to represent the possibility of unknown hemorrhage events. For this purpose, we consider a stochastic model of the following form:

$$[J_H]_1 = \epsilon_g,\ \epsilon_g \sim \mathcal{P}(0, \sigma_H, \xi_H) \quad (29)$$

$$[J_H]_k = (1 - \epsilon_b)[J_H]_{k-1} + \epsilon_b \epsilon_g, \epsilon_b \sim \mathcal{B}(1, \gamma_H),\ \epsilon_g \sim \mathcal{P}(0, \sigma_H, \xi_H) \quad (30)$$

In this model, first, possible hemorrhage rates $[J_H]_1$ are instantiated in (29) by drawing samples ϵ_g from a generalized Pareto distribution (denoted by $\mathcal{P}$) located at zero, where σ_H is the scale parameter and ξ_H is the shape parameter. This distribution represents a range of possible hemorrhage rates, where lower hemorrhages have a higher probability of occurrence. Over time, the rate of hemorrhage in a patient follows the relationship in (30), where ϵ_b is a sample from the Bernoulli distribution, and the probability γ_H acts as a forgetting factor for the hemorrhage rate. According to this model, the rate of hemorrhage in a patient retains its value over short periods of time, except for the possibility of transitioning to a different rate drawn from the hemorrhage model in (29). Overall, this model is built to inform a PIPF-based monitoring algorithm about the possibility of unknown hemorrhage events, and furthermore allow the algorithm to form beliefs about the presence of unknown hemorrhage in a patient over time.

4) Physiological Measurement Model: At the third level in the hierarchical model in Section II.A, a physiological measurement model $\mathcal{M}$ generates observations from the characteristics and state evolutions generated by the first two levels (i.e., $\mathcal{G}$ and $\mathcal{H}$). In the context of hypovolemia treatment, mean arterial pressure measurements are typically available in clinical settings, while cardiac output and hematocrit measurements may be available only in experimental settings. To represent these potentially measured physiological variables, we consider the following model:

$$[y_{HCT}]_k = [H]_k + n_{HCT}\epsilon, \epsilon \sim \mathcal{N}(0, I) \quad (31)$$

$$[y_{CO}]_k = [Q]_k + n_{CO}\epsilon, \epsilon \sim \mathcal{N}(0, I) \quad (32)$$

$$[y_{MAP}]_k = [P_a]_k + n_{MAP}\epsilon, \epsilon \sim \mathcal{N}(0, I) \quad (33)$$

where y_{HCT}, y_{CO}, and y_{MAP} denote the generated measurements for hematocrit, cardiac output, and mean arterial pressure, respectively, and n_{HCT}, n_{CO}, and n_{MAP} denote the standard deviation of noise acting on these measurements. Overall, this measurement model is intended to inform a PIPF-based monitoring algorithm about the noises/artifacts that may corrupt each measured variable, thereby giving the algorithm a way to attribute such disturbances as they occur.

B. The PIPF-Based Monitoring Algorithm

As suggested in Section II, given a generative model of the relevant physiological phenomena (such as the model introduced in Section III.A), Algorithm 1 can be used to create a monitoring system for hypovolemia treatment. The notable steps of this procedure are the following:

- To perform $\theta_1^i \sim \mathcal{G}_1(\phi)$, we use (17) to generate N proposed samples θ_1^i representing a patient's possible baseline characteristics.
- To perform $x_1^i \sim \mathcal{H}_1(\theta_1^i)$, we use the generated θ_1^i's to initialize the states of the physiological model in (20)-(22), (24), (26), (28) as follows: v_a is set to $v_{a0} = 0.3v_0$, v_v is set to $v_{v0} = 0.7v_0$, v_r is set to $v_{r0} = H_0 v_0$, while r_F, s_R, and s_Q are set to zero. Also, the possible hemorrhage rates are initialized according to (29).
- To perform $\theta_k^i \sim \mathcal{G}_k(\theta_{k-1}^i, \phi)$, we use the relationship in (18).
- To perform $x_k^i \sim \mathcal{H}_k(x_{k-1}^i, \theta_{k-1}^i, u_{k-1})$, we use the relationships in (20)-(28), (30).
- To perform $\log \alpha^i \leftarrow \log p(y_k | x_k^i, \theta_k^i, n)$, we consider the measurement model in (31)-(33), which implies a likelihood function of the following form:

$$\log p\left(y_k | x_k^i, \theta_k^i, n\right) = \sum_m \left[-\frac{1}{2n_m^2}\left([y_m]_k - [y_m]_k^d\right)^2 - \log(n_m) - \frac{1}{2}\log(2\pi) \right] \quad (34)$$

where the index m enumerates the available measured variables at time k (e.g., if mean arterial pressure and hematocrit measurements are available at time k, then $m \in \{MAP, HCT\}$), and $[y_m]_k^d$ denotes the measured value. Executing Algorithm 1 as outlined above results in a monitoring algorithm that takes in measurements as they become available over time, forming running beliefs about the physiological states (i.e., v_a, v_v, v_r, r_F, s_R, and s_Q), characteristics (i.e., shown in the θ_k vector in (19)) and events (i.e., J_H) in a patient.

C. Physiological Data

To demonstrate the potential merits and limitations of the PIPF-based monitoring algorithm, we utilize a physiological dataset from previous work [33]–[35], which contains time-series measurements pertaining to hypovolemia treatment in 23 animal (sheep) experiments. In each experiment, the animal is subjected to controlled hemorrhage and subsequently resuscitated over time using fluid (crystalloid) infusions, which are administered according to the recommendations of a control algorithm. Each experiment spans a course of 180 minutes, where the subject's hematocrit, cardiac output, and mean arterial pressure are measured approximately every 5 minutes. This dataset is especially suited to the purpose of analyzing the PIPF-based monitoring algorithm, as it can be used to assess the algorithm's ability to form reliable beliefs about variables that are often unmeasured or unknown in clinical settings (e.g., hematocrit, cardiac output, and hemorrhage rates) by processing measurements that are typically available in clinical settings (e.g., mean arterial pressure and infusion rates). In the next section, we provide an overview of our data analysis procedures for this purpose.

D. Data Analysis

1) Problem Setting: To evaluate the PIPF algorithm in a scenario that resembles real-world hypovolemia treatment, we consider the task of monitoring each subject in our dataset as it undergoes controlled hemorrhage and fluid resuscitation. In this scenario, the monitoring algorithm receives a stream of mean arterial pressure and infusion rate measurements (which are typically available in clinical settings) and is tasked with forming and updating beliefs about the subject's states, characteristics, and hemorrhage events over time. These beliefs are in turn evaluated against available measurements from the subject not shown to the monitoring algorithm (i.e., hematocrit, cardiac output, and hemorrhage rates) in order to assess its performance. It is important to note that this problem setting is a purposefully challenging one, where many unmeasured variables are inferred from few measured variables. This setting enables us to assess the merits and limitations of the PIPF-based monitoring algorithm especially when the clinically available data are limited.

2) Algorithm Evaluation: As presented in Section II.A, obtaining a PIPF-based monitoring algorithm involves (i) training a generative physiological model on a physiological dataset obtained from the population, and (ii) providing this model to Algorithm 1. To evaluate the performance of the monitoring algorithm, we follow a leave-one-out cross-validation procedure. For each studied subject, we exclude the subject from the population dataset used to train the generative physiological model and use the resulting model in Algorithm 1 to obtain a PIPF-based monitoring algorithm. Then, we test the resulting algorithm on the excluded subject by providing the subject's mean arterial pressure and infusion rate data to the algorithm as a stream of measurements and assessing the performance of the algorithm based on the adequacy of its beliefs about the subject's unseen measurements (i.e., hematocrit, cardiac output, and hemorrhage rates).

To quantify the adequacy of the beliefs provided by the algorithm, we utilize the mean continuous ranked probability score (MCRPS), which is a metric suitable for comparing beliefs to measured values [36]. A particle-based form of this metric can be written as follows:

$$\text{MCRPS}(F_m, y_m) = \mathbb{E}_k \left[\mathbb{E}_{z_1 \sim [F_m]_k} |z_1 - [y_m]_k| - \frac{1}{2}\mathbb{E}_{z_1, z_2 \sim [F_m]_k} |z_1 - z_2| \right] \quad (35)$$

where $[F_m]_k$ denotes the belief provided by the monitoring algorithm about variable m at time k, which is expressed as a collection of weighted points, z_1 and z_2 are samples from this belief, and $[y_m]_k$ denotes the measured value for variable m at time k. Intuitively, this scoring function maximally promotes predictions that are sharp (i.e., certain) and accurate, and maximally discourages predictions that are sharp and inaccurate. Smaller MCRPS scores indicate better predictions and, in the case of deterministic predictions, the MCRPS metric reduces to mean absolute error.

3) PIPF versus Particle Filtering (PF): To highlight the potential advantages of the PIPF-based monitoring algorithm with respect to a more established approach that does not consider population-level information, we build an alternative monitoring algorithm that operates based on the traditional particle filtering (PF) scheme [14], [21]. This PF-based monitoring algorithm differs from its PIPF-based counterpart in that it does not include a patient generator model in its underlying model. According to PF convention, the unmeasured variables of the problem (i.e., the subject's states, characteristics, and events) are instead initialized by drawing a uniform sample of particles over a plausible range in the space

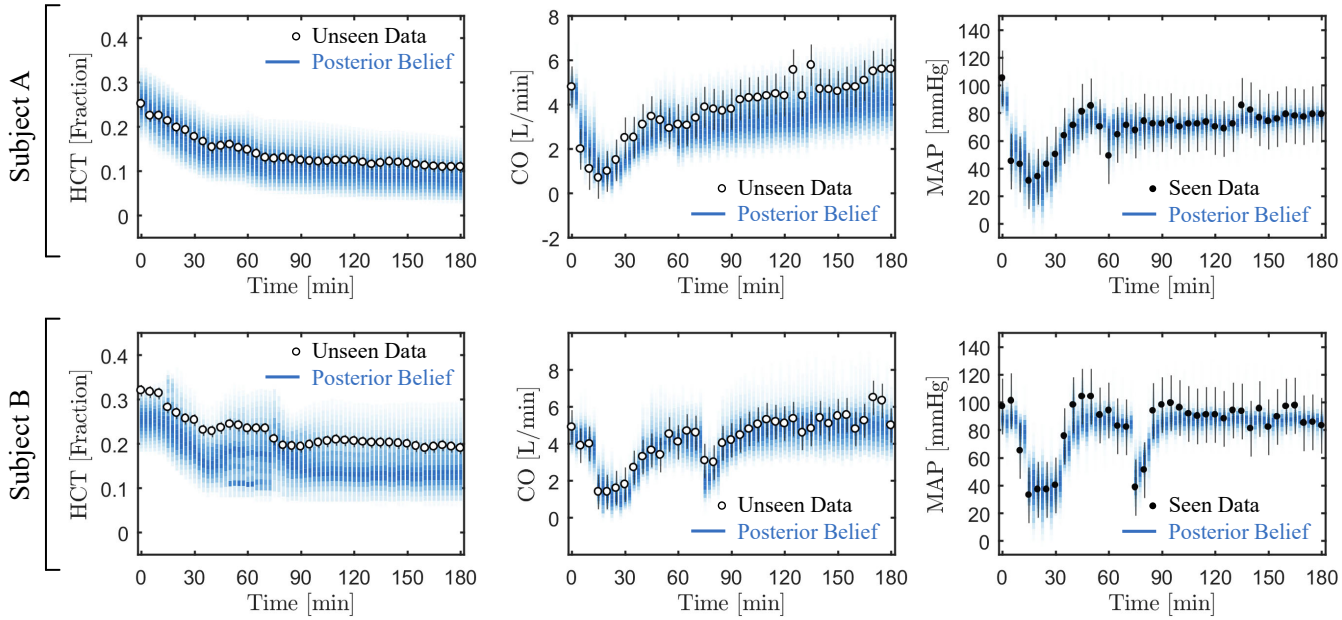

Fig. 2. Marginalized posterior beliefs about cardiac output (CO) and hematocrit (HCT) in two representative subjects when presenting the PIPF algorithm with a sequence of mean arterial pressure (MAP) measurements. Bolder colors represent higher belief density. Black dots show data presented to PIPF, while white dots show data never presented to PIPF.

of unmeasured variables. In addition, to enable the PF-based monitoring algorithm to adapt to the subject's characteristics, we consider a random walk model for the parameters over time $\theta_k = \theta_{k-1} + \sigma_{PF}\epsilon$ where $\epsilon \sim \mathcal{N}(0, I)$. To compare the PIPF-based versus PF-based monitoring algorithms, we use the MCRPS scoring procedure described above. To test the significance of the score differences between the two approaches, we use the Wilcoxon signed-rank test. Overall, this comparison study is designed primarily to highlight the potential merits and limitations of incorporating population-level information into the filtering process.

IV. Results and Discussion

Forming reliable beliefs about a patient's state is essential for algorithmic decision-making in medical care settings. This task is especially challenging when the physiological measurements available from the patient are intermittent or contain limited information. To address this challenge, we proposed the PIPF scheme, where the information encoded in a generative physiological model is leveraged to form more robust beliefs about a patient's state. This section presents the results of applying the PIPF scheme to the problem of monitoring for hypovolemia treatment and discusses the merits and limitations of the PIPF scheme in this context.

A. Beliefs about Unmeasured Physiological Variables

Fig. 2 shows the beliefs formed by the PIPF-based monitoring algorithm about hematocrit (HCT), cardiac output (CO), and mean arterial pressure (MAP) in two representative subjects (marked by Subject A and Subject B). These beliefs were formed when providing the algorithm with a stream of MAP (see in Fig. 2; right column) and infusion rate (see Fig. 3; blue line) measurements. Fig. 5 shows (in the bottom panel) the beliefs formed about the internal states and characteristics of Subject A in this scenario. As it is expected from the formulation of the PIPF algorithm, whenever a new measurement becomes available, the algorithm adjusts its beliefs about the subject's physiological states and characteristics. As a result, beliefs about the subject's MAP closely follow the MAP data, and the algorithm can provide beliefs about the subject's HCT and CO (see Fig. 2) as well as its internal states and characteristics (see Fig. 5). These beliefs appear consistent with the HCT and CO data unseen by the algorithm.

These representative results also highlight two notable behavior patterns in the beliefs generated by the PIPF algorithm in relation to the informativeness of the data. First, CO is an example of a variable that could be *strongly informed* by MAP data because of its close physiological relationship with MAP. Namely, the pressure in the arterial space, which has a relatively low compliance, is strongly affected by the flow rate of blood coming into the space. As a result, in the initial stages of the experiment, where variations in hemorrhage and infusion protocols (e.g., see Fig. 3) are expected to excite the subject's physiological dynamics, the monitoring algorithm can infer relatively sharp beliefs about CO that are also consistent with the corresponding CO data (see Fig. 2; center column). Second, baseline HCT is an example of a variable that is *weakly informed* by MAP data because of its distant physiological relationship with MAP. Namely, MAP is primarily affected by blood volume kinetics (i.e., the volume of blood that ends up in the arterial space at each time) while baseline HCT is a measure of blood composition. In this scenario, the monitoring algorithm forms its beliefs about baseline HCT in large part from the population-level information encoded in the generative physiological model. As a result, the beliefs about baseline HCT appear to span the range of plausible HCT values in the population (see Fig. 2; left column). Overall, these results suggest that the PIPF-based monitoring algorithm shows

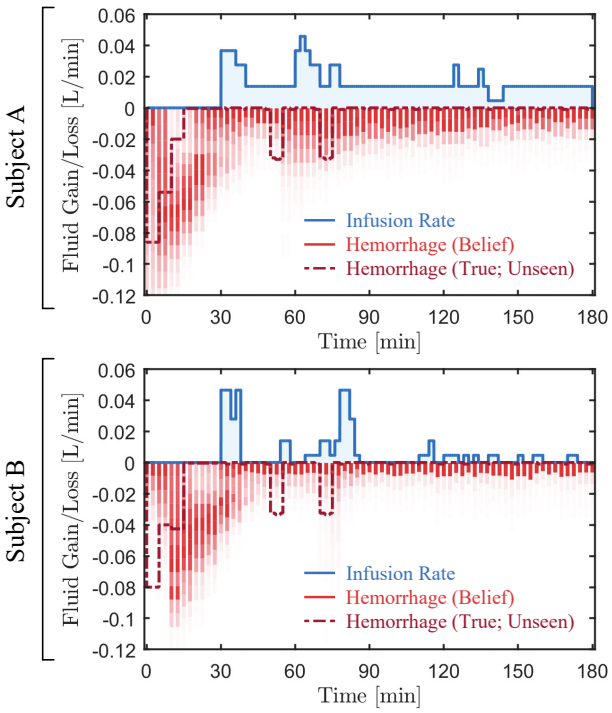

Fig. 3. Marginalized posterior beliefs about hemorrhage rate in two representative subjects when presenting the PIPF algorithm with a sequence of mean arterial pressure (MAP) measurements. Bolder colors represent higher belief density. True hemorrhage rates (unknown to the algorithm) and infusion rates (known to the algorithm) are provided for reference.

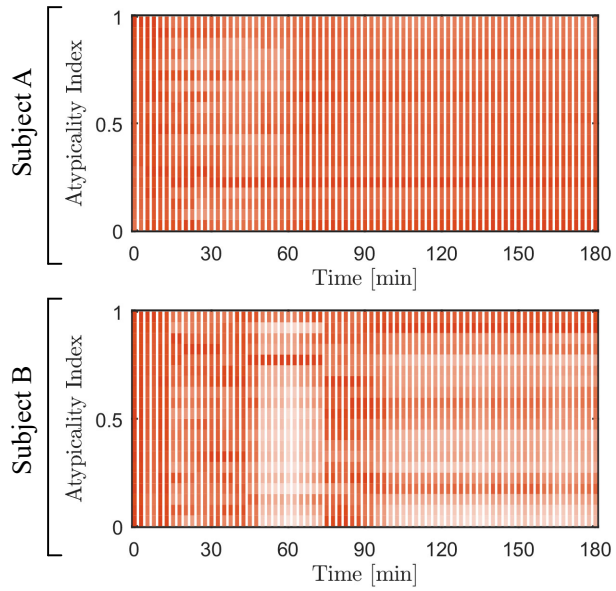

Fig. 4. Marginalized posterior beliefs about atypicality in two representative subjects when presenting the PIPF algorithm with a sequence of mean arterial pressure (MAP) measurements. Bolder colors represent higher belief density.

promise in forming beliefs about the unmeasured variables of the physiological system by combining the information contained in subject-specific measurements with the population-level information encoded in the generative physiological model.

B. Beliefs about Subject Characteristics

Fig. 5 shows (in the top panel) the posterior beliefs about the characteristics of Subject A. These beliefs were formed when providing the algorithm with a stream of MAP (see in Fig. 2; top-right column) and infusion rate (see Fig. 3; top-panel; blue line) measurements. For most characteristics, the beliefs maintain a high entropy (i.e., large spread) throughout the filtering process, and resemble the characteristics generated by the generative model. However, in the initial phases of the experiment (e.g., first 60 minutes), where the subject undergoes significant perturbations (i.e., hemorrhage and fluid infusions), the beliefs show some deviations from the characteristics generated by the generative model. The high entropy of the beliefs indicates that the characteristics of the subject are, for the most part, only weakly identifiable from a stream of MAP and infusion rate measurements, especially when the perturbations do not sufficiently excite the subject's physiology. By design, the PIPF algorithm reverts to the characteristics generated by the generative physiological model in such conditions. In other words, in the absence of informative measurements or excitations, the algorithm opts for considering all possibilities with regards to subject characteristics as informed by the generative physiological model.

Despite the high entropy, the beliefs about subject characteristics may be used to calculate useful summarized information about the subject. The *subject atypicality index* [37] is a notable example of this, which measures the presence of atypical characteristics in a subject by comparing the subject's characteristics to those of the population. Fig. 4 shows the posterior beliefs about atypicality in two representative subjects. The index values lie between zero and one, with a value of one indicating a highly atypical subject. For Subject A, the beliefs about atypicality show a high level of entropy and uniformly span the range between zero and one, which indicates that Subject A's atypicality cannot be established or rejected based on the given stream of MAP and infusion rate data. For Subject B, the beliefs about atypicality shift toward the higher end of the spectrum at two times during the experiment (starting approximately at the 50-minute mark for 20 minutes and at the 95-minute mark for the rest of the experiment). Inspecting Subject B's MAP data at those times reveals two instances of rapid rise in pressure to values even higher than the subject's baseline (pre-hemorrhage) pressure, which is not generally expected from a typical subject undergoing hemorrhage and crystalloid resuscitation. Overall, these results suggest that, despite the high entropy, beliefs about a subject's characteristics may be summarized to obtain potentially useful information about the subject.

C. Beliefs about Physiological Events

Fig. 3 shows beliefs about hemorrhage rate in two representative subjects when providing the monitoring algorithm with a stream of MAP and infusion rate measurements. In Subject A, the algorithm attributes the subject's sudden initial drop in MAP (see Fig. 2; top-right panel) to the presence of hemorrhage, with beliefs about hemorrhage rate that are consistent with hemorrhage data unseen by the algorithm. Subsequently, the algorithm infers a

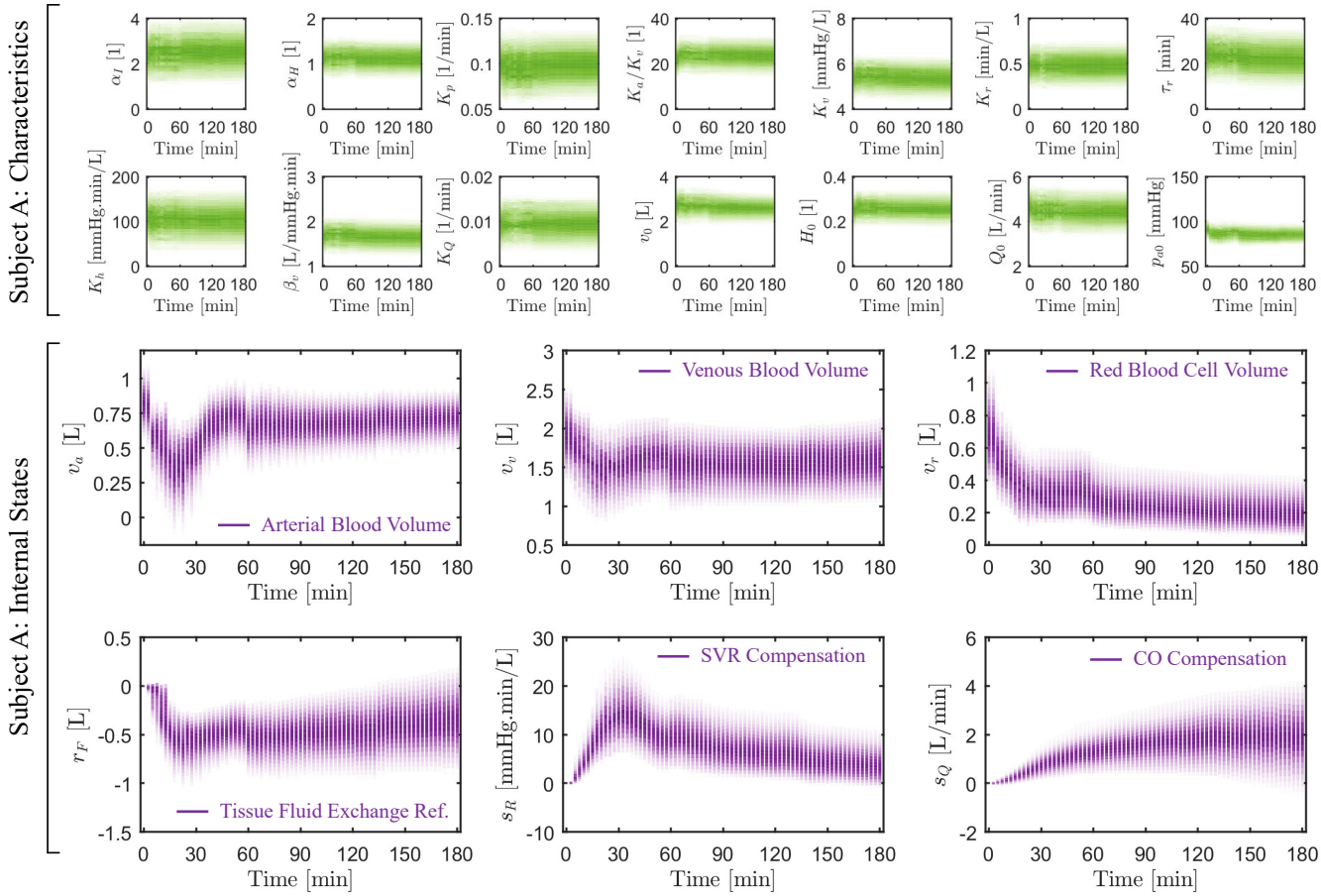

Fig. 5. Marginalized posterior beliefs about a representative subject's states and characteristics when presenting the PIPF algorithm with a sequence of mean arterial pressure (MAP) measurements. Bolder colors represent higher belief density.

Table I. Comparing cross-validation scores (based on MCRPS) between the PIPF algorithm and a conventional particle filtering (PF) approach, when presenting the algorithms with a sequence of mean arterial pressure (MAP) measurements [Median (Q1, Q3); N=23; Lower is better]. *: p<0.05.

Method	Hemorrhage Rate [mL/min]	Hematocrit [%]	Cardiac Output [L/min]	Mean Arterial Pressure [mmHg]
PIPF	10.03* (8.29, 11.10)	2.15* (1.25, 3.91)	0.52 (0.35, 0.89)	3.85 (3.70, 4.11)
PF	20.29 (17.17, 22.37)	4.97 (4.18, 6.54)	0.53 (0.42, 0.97)	3.81 (3.67, 3.89)

cessation of hemorrhage (albeit with delay) and provides reasonable beliefs about the later instances of small hemorrhage. In Subject B, the algorithm detects the initial hemorrhage, but with some delay, which can be attributed to the first few MAP measurements remaining high despite the presence of large hemorrhage (see Fig. 2; bottom-right panel). Subsequently, the algorithm detects one of the two smaller hemorrhages, depending on whether they affect the MAP measurements. Overall, these results suggest that the PIPF-based monitoring algorithm may be utilized to form useful beliefs about physiological events in a subject (such as hemorrhage), when the underlying model includes components that represent the event, and the occurrence of the event leaves detectable effects on the measurements available from the subject.

D. Effect of Population-Level Information

Table I compares cross-validation scores (based on MCRPS in N=23 subjects) between the PIPF algorithm and a conventional particle filtering (PF) algorithm described in Section III.D.3. Comparing the scores between the two cases reveals that the PIPF approach provides superior beliefs about hemorrhage rate and HCT when compared to the conventional PF algorithm, while beliefs about CO are comparable between the two cases. The advantage of the PIPF algorithm may be explained by inspecting the structure of the inference problem that underlies its operation (Fig. 1(b)). In this structure, the latent variables of the problem (i.e., $\theta_{1:k}$, $x_{1:k}$) are informed by a patient generator model (parameterized by ϕ) that is derived from past population-level data. In contrast, the conventional PF algorithm relies on initialization and transition procedures for $\theta_{1:k}$, $x_{1:k}$ that do not leverage past data (see Section III.D.3). As a result, the PIPF algorithm tends to exhibit superior performance, especially for those latent variables that are weakly informed by the available measurements. As presented in Section IV.A, CO is a variable that is expected to be strongly informed by MAP data due to its close physiological relationship with MAP. As a result, both algorithms can infer CO in large part from the information contained in MAP, giving relatively sharp beliefs about CO that are also consistent with the CO data. This results in comparable MCRPS scores for the

two algorithms. In contrast, hemorrhage rate and HCT are variables that are expected to be weakly informed by MAP data because of their more distant physiological relationship with MAP. As a result, the PIPF algorithm forms its beliefs about hemorrhage rate and HCT in large part from the population-level information encoded in the generative physiological model, while the PF algorithm does not have access to the population-level information. This results in superior MCRPS scores for the PIPF-based monitoring algorithm. Overall, these results suggest that incorporating population-level information into the filtering process (as it is done in the PIPF algorithm) results in the formation of superior beliefs, especially for physiological variables that are weakly informed by the available subject-specific data.

E. Potential Applications

As presented in Section IV-A through C, the PIPF approach generates real-time beliefs (in the form of probability distributions) about a patient's unmeasured physiological variables, characteristics, and events, given a stream of physiological measurements. These beliefs can in turn be passed on to human users, decision-support algorithms, and/or closed-loop control algorithms as a basis for decision-making. The following paragraphs briefly discuss these use cases.

Utility for Users and/or Clinicians: In order to facilitate the human user's use and interpretation of the beliefs generated by PIPF, the corresponding distributions may be converted into point-estimates (e.g., by taking the central tendency of the distribution) and credible intervals (e.g., by taking the 10-90th percentiles of the distribution). In this way, the user would be equipped with, respectively, a "best estimate" and a "confidence measure" for each variable of interest, both of which can be informative in the course of decision-making. To illustrate this aspect, the supplementary material accompanying this paper includes example visualizations (Fig. S1, Fig. S2, and Fig. S3) and performance metrics (Table S-I and Table S-II) for this human-friendly representation.

Utility for Decision/Control Algorithms: A main advantage of PIPF lies in the fact that it uses collections of samples (i.e., particles) to represent beliefs. As a result, it would be possible for a PIPF-based monitoring algorithm to form highly expressive beliefs (i.e., beliefs with complex shape; e.g., asymmetric and/or multi-modal) if necessary. These sample-based beliefs can therefore act as rich, real-time representations of likely values and uncertainties associated with a patient's unmeasured physiological variables. Arguably, decision algorithms could be designed to perform principled risk/reward analysis on these beliefs in order to suggest/apply best courses of action for a given patient. We believe this aspect to be a worthwhile avenue for future work.

V. Conclusion

In this work, we proposed the population-informed particle filter (PIPF), a Bayesian filtering approach that leverages a generative physiological model to provide beliefs about a patient's states, characteristics, and events in the context of physiological monitoring. Using a case study on monitoring for hemodynamic management, we showed that the PIPF approach can provide reasonable beliefs (as compared to excluded data) about the likely values and uncertainties associated with a patient's physiological variables (e.g., hematocrit and cardiac output), characteristics (e.g., tendency for atypical behavior), and events (e.g., hemorrhage). In addition, we demonstrated that incorporating population-level information into the filtering process (as is done in the PIPF algorithm) results in the formation of beliefs that are superior to those provided by a traditional particle filtering approach, especially for physiological variables that are weakly informed by the available patient-specific measurements. These results imply that PIPF is a promising candidate for use in physiological monitoring systems that are required to form beliefs about unmeasured aspects of a patient's physiology by processing low-information and intermittent physiological measurements. Therefore, future efforts should be devoted to applying and evaluating the PIPF approach in a wider range of physiological monitoring applications, and researching principled ways to maximally leverage the beliefs provided by PIPF to design next-generation physiological decision-support and closed-loop control algorithms.

Acknowledgement

The authors would like to thank Dr. George C. Kramer at the University of Texas Medical Branch for sharing the in vivo animal dataset.

References

[1] J. F. Dyro, *Clinical engineering handbook*. Elsevier Science Ltd., 2004.

[2] M. N. Sawka and K. E. Friedl, "Emerging Wearable Physiological Monitoring Technologies and Decision Aids for Health and Performance," *Journal of Applied Physiology*, vol. 124, no. 2, pp. 430–431, 2018.

[3] J.-O. Hahn and O. T. Inan, "Physiological closed-loop control in critical care: Opportunities for innovations," *Progress in Biomedical Engineering*, May 2022.

[4] C. Zaouter, A. Joosten, J. Rinehart, M. M. R. F. Struys, and T. M. Hemmerling, "Autonomous Systems in Anesthesia: Where Do We Stand in 2020? A Narrative Review," *Anesth Analg*, vol. 130, no. 5, pp. 1120–1132, 2020.

[5] E. Brogi, S. Cyr, R. Kazan, F. Giunta, and T. M. Hemmerling, "Clinical Performance and Safety of Closed-Loop Systems: A Systematic Review and Meta-analysis of Randomized Controlled Trials," *Anesth Analg*, vol. 124, no. 2, pp. 446–455, Feb. 2017.

[6] A. Weisman, J. W. Bai, M. Cardinez, C. K. Kramer, and B. A. Perkins, "Effect of artificial pancreas systems on glycaemic control in patients with type 1 diabetes: a systematic review and meta-analysis of outpatient randomised controlled trials," *Lancet Diabetes Endocrinol*, vol. 5, no. 7, pp. 501–512, Jul. 2017.

[7] A. Gelb and others, *Applied optimal estimation*. MIT press, 1974.

[8] B. D. O. Anderson and J. B. Moore, *Optimal filtering*. Courier Corporation, 2012.

[9] B. D. O. Anderson and J. B. Moore, "Kalman Filtering: Whence, What and Whither?," *Mathematical System Theory*, pp. 41–54, 1991.

[10] J. Humpherys, P. Redd, and J. West, "A fresh look at the kalman filter," *SIAM Review*, vol. 54, no. 4, pp. 801–823, 2012.

[11] E. A. Wan and R. van der Merwe, "The unscented Kalman filter for nonlinear estimation," *IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium, AS-SPCC 2000*, pp. 153–158, 2000.

[12] H. M. T. Menegaz, J. Y. Ishihara, G. A. Borges, and A. N. Vargas, "A Systematization of the Unscented Kalman Filter Theory," *IEEE Transactions on Automatic Control*, vol. 60, no. 10, pp. 2583–2598, Oct. 2015.

[13] G. A. Einicke and L. B. White, "Robust extended Kalman filtering," *IEEE Transactions on Signal Processing*, vol. 47, no. 9, pp. 2596–2599, 1999.

[14] S. Särkkä, *Bayesian filtering and smoothing*, no. 3. Cambridge university press, 2013.

[15] R. van de Schoot *et al.*, "Bayesian statistics and modelling," *Nature Reviews Methods Primers 2021 1:1*, vol. 1, no. 1, pp. 1–26, Jan. 2021.

[16] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*. Chapman and Hall/CRC, 1995.

[17] O. Cappe, S. J. Godsill, and E. Moulines, "An overview of existing methods and recent advances in sequential Monte Carlo," *Proceedings of the IEEE*, vol. 95, no. 5, pp. 899–924, 2007.

[18] H. R. Künsch, "Particle filters," *Bernoulli*, vol. 19, no. 4, pp. 1391–1403, Sep. 2013.

[19] N. Kantas, A. Doucet, S. S. Singh, J. Maciejowski, and N. Chopin, "On particle methods for parameter estimation in state-space models," *Statistical science*, vol. 30, no. 3, pp. 328–351, 2015.

[20] C. A. Naesseth, F. Lindsten, and T. B. Schön, "Elements of Sequential Monte Carlo," *Foundations and Trends® in Machine Learning*, vol. 12, no. 3, pp. 307–392, Nov. 2019.

[21] A. Doucet and A. M. Johansen, "A Tutorial on Particle Filtering and Smoothing: Fifteen Years Later," *Oxford Handbook of Nonlinear Filtering*, vol. 12, pp. 656–704, 2011.

[22] V. Šmídl and A. Quinn, "Variational Bayesian filtering," *IEEE Transactions on Signal Processing*, vol. 56, no. 10 II, pp. 5020–5030, 2008.

[23] K. J. Friston, "Variational filtering," *Neuroimage*, vol. 41, no. 3, pp. 747–766, Jul. 2008.

[24] J. Marino, M. Cvitkovic, and Y. Yue, "A General Method for Amortizing Variational Filtering," in *Advances in Neural Information Processing Systems*, 2018, vol. 31.

[25] C. Naesseth, S. Linderman, R. Ranganath, and D. Blei, "Variational Sequential Monte Carlo," in *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, May 2018, vol. 84, pp. 968–977.

[26] A. Tivay, G. C. Kramer, and J. O. Hahn, "Collective Variational Inference for Personalized and Generative Physiological Modeling: A Case Study on Hemorrhage Resuscitation," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 2, pp. 666–677, Feb. 2022.

[27] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational Inference: A Review for Statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518. American Statistical Association, pp. 859–877, Apr. 03, 2017.

[28] P. Blanchard, D. J. Higham, and N. J. Higham, "Accurately computing the log-sum-exp and softmax functions," *IMA Journal of Numerical Analysis*, vol. 41, no. 4, pp. 2311–2330, Oct. 2021.

[29] A. E. Laher *et al.*, "A review of hemodynamic monitoring techniques, methods and devices for the emergency physician," *The American Journal of Emergency Medicine*, vol. 35, no. 9, pp. 1335–1347, Sep. 2017.

[30] B. Saugel, J. L. Vincent, and J. Y. Wagner, "Personalized hemodynamic management," *Current Opinion in Critical Care*, vol. 23, no. 4, pp. 334–341, Aug. 2017.

[31] A. Bougié, A. Harrois, and J. Duranteau, "Resuscitative strategies in traumatic hemorrhagic shock," *Annals of Intensive Care*, vol. 3, no. 1, pp. 1–9, Jan. 2013.

[32] A. Tivay, G. C. Kramer, and J. O. Hahn, "Virtual Patient Generation using Physiological Models through a Compressed Latent Parameterization," *Proceedings of the American Control Conference*, vol. 2020-July, pp. 1335–1340, Jul. 2020.

[33] N. R. Marques *et al.*, "Automated closed-loop resuscitation of multiple hemorrhages: a comparison between fuzzy logic and decision table controllers in a sheep model," *Disaster Mil Med*, vol. 3, no. 1, Dec. 2017.

[34] S. U. Vaid *et al.*, "Normotensive and hypotensive closed-loop resuscitation using 3.0% NaCl to treat multiple hemorrhages in sheep," *Crit Care Med*, vol. 34, no. 4, pp. 1185–1192, Apr. 2006.

[35] A. D. Rafie *et al.*, "Hypotensive resuscitation of multiple hemorrhages using crystalloid and colloids," *Shock*, vol. 22, no. 3, pp. 262–269, Sep. 2004.

[36] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *J Am Stat Assoc*, vol. 102, no. 477, pp. 359–378, 2007.

[37] A. Tivay, G. C. Kramer, and J. O. Hahn, "Inference-based subject atypicality and signal quality indicators for physiological data," *MCPS 2021 - Proceedings of the 2021 Medical Cyber Physical Systems and Internet of Medical Things*, pp. 7–11, May 2021.