doi:10.11959/j.issn.1000-436x.2015247

传感器网络环境监测时间序列数据的高斯过程建模与多步预测

陈艳[1,2]，王子健[1]，赵泽[1]，李栋[1]，崔莉[1]
（1. 中国科学院 计算技术研究所，北京 100190；2.中国科学院大学，北京 100190）

摘 要：针对传感网环境监测应用采集的时间序列数据，提出了一种新的基于高斯过程模型的多步预测方法，实现了对未来时刻的环境监测数据的预测。高斯过程模型通过核函数描述数据的特性，通过对环境监测数据的经验模态分解，以及对其内在物理特性的分析，构建了针对环境监测数据的高斯过程核函数，实现了对数据变化模式的描述。在基于 3 个数据集的 5 个种类、20 000 多个环境监测数据上进行了性能对比实验，结果表明，与对比预测方法相比，提出的高斯过程多步预测方法对未来时刻的环境监测数据的平均预测精度可以提高 20%，可以应用于环境参数未来趋势分析、异常环境事件预警等场景。

关键词：传感网环境监测；时间序列；高斯过程；多步预测
中图分类号：TP393 **文献标识码**：A

Gaussian process modeling and multi-step prediction for time series data in wireless sensor network environmental monitoring

CHEN Yan[1,2], WANG Zi-jian[1], ZHAO Ze[1], LI Dong[1], CUI Li[1]
(1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;
2. University of Chinese Academy of Sciences, Beijing 100190, China)

Abstract: For time series data collected from WSN environmental monitoring applications, a novel multi-step prediction method based on Gaussian process model was proposed. The method could make prediction for future environmental monitoring data. Kernel functions were used to describe data properties in the Gaussian process model. Kernel functions for environmental monitoring data were constructed through the EMD (empirical mode decomposition) technique and analysis of data inherent physical properties. And the constructed kernel functions were capable of describing the data change mode. Extensive experiments for multi-step prediction performance comparison test were performed on three kinds of data sets using over 20 000 environmental monitoring data records. Experimental results show that the average prediction accuracy of the Gaussian process multi-step prediction method can be increased by 20% than compared prediction methods. The prediction method can be applied to future environmental parameters trend analysis, early warning for abnormal environmental events and other scenes.

Key words: WSN environmental monitoring; time serie; Gaussian process; multi-step prediction

1 引言

无线传感器网络(WSN, wireless sensor network)[1]以其低功耗、低成本、分布式和自组织的特点在环境监测领域得到了广泛的应用[2]。在环境监测应用中，传感器节点部署于用户关心的环境中（如湖泊、展陈室等），定期采集用户感兴趣的环境数据（如 PH、温度等），并上传到数据中心进行存储和显示。目前，这些采集到的大量环境监测数据并未得到有效的利用，通常仅用于历史数据的查询和实时数据的显示。

事实上，采集到的环境监测数据按照时间顺序

收稿日期：2014-11-06；**修回日期**：2015-03-08
基金项目：中国科学院战略性先导科技专项基金资助项目（XDA06010403）；国家国际科技合作专项基金资助项目（2013DFA10690）；国家自然科学基金资助项目(61100179, 61202412)
Foundation Items: The Strategic Priority Research Program of the Chinese Academy of Sciences (XDA06010403); The International S&T Cooperation Program of China (ISTCP) (2013DFA10690);The National Natural Science Foundation of China (61100179, 61202412)

组织起来可以形成一个时间序列[3]，由于环境参数的内在物理特性，该时间序列通常具有特定的变化模式。如果能够有效利用大量的环境数据，发现其内在变化模式，进而预测环境数据的变化趋势，则可以及早发现环境异常事件，便于用户及时采取有效的应急措施，这对环境监测领域具有重要的应用价值[4]。例如，在一个湖泊生态监测系统中，当水体的PH、溶解氧等水质参数发生异常变化时，可能标识着蓝藻水华的发生。如果能够提前预测水质参数的变化趋势，及早发现异常变化，就可以进行蓝藻水华的预警并提前制定应急方案，降低其带来的损失。

时间序列建模与多步预测方法可以用于建立时间序列数据的模型，并对未来时刻的数据进行预测。对于在时刻 T 之前采集的、采样间隔为 W 的 M 个环境监测时间序列 $S=\{s_M,s_{M-1},\cdots,s_1\}$ 数据（s_i 为第 i 个环境监测时间序列数据，$i=1,2,\cdots,M$），多步预测会估计 H 步之后的数据取值，即预测 $T+WH$ 时刻的数据 s_{M+H}，进而得到环境数据的未来变化趋势。

本文应用高斯过程(GP, Gaussian process)建立环境监测时间序列数据的模型，基于该模型提出了多步预测方法，即首先对环境监测数据进行经验模态分解，然后分析数据物理特性（如周期性和局部动态变化特性），进而构建环境监测数据的高斯过程核函数来描述数据特性，该模型能够对未来时刻的环境监测数据进行预测。针对现有实际部署系统的多种环境监测数据，进行了大量的预测精度性能对比实验。实验结果表明，在所有数据集上，构建的高斯过程多步预测方法的平均预测精度较之对比的预测方法可以提高 20%。

2　相关工作与背景知识

时间序列建模和多步预测涉及到 2 个方面，即预测策略和预测模型。

2.1　预测策略

在预测策略方面，现阶段应用的预测策略主要有迭代策略、直接策略和多输入—多输出（MIMO, multi-input and multi-output）策略。

2.1.1　迭代策略

这种策略是通过最小化单步预测误差的平方和来训练一个预测模型，同时以得到的预测值作为该模型的输入，来得到下一个预测值，直到达到多步预测的步长，Chevillon 称这种策略为迭代策略[5]。

迭代策略首先将一维原始时间序列数据组织成一个输入—输出的数据格式

$$D=\left\{(\boldsymbol{x}_t,y_t)\in\left(\mathbf{R}^P\times\mathbf{R}\right)\right\}^N \tag{1}$$

其中，$\boldsymbol{x}_t=\{s_t,\cdots,s_{t-P+1}\}^P$，$y_t=s_{t+1}$，$P$ 为回归阶数。

然后开始训练一个单步预测模型

$$y_t=f(\boldsymbol{x}_t)+\omega \tag{2}$$

其中，ω 表示噪音。

基于训练模型，计算接下来的 H 个数据值

$$\hat{s}_{t+h}=\begin{cases}\hat{f}\left(s_t,s_{t-1},\cdots,s_{t-P+1}\right) & ,h=1\\ \hat{f}\left(\hat{s}_{t+h-1},\cdots,\hat{s}_{t+1},s_t,\cdots,s_{t-P+h}\right), & h\in[2,\cdots,P]\\ \hat{f}\left(\hat{s}_{t+h-1},\cdots,\hat{s}_{t+h-P}\right) & ,h\in[P+1,\cdots,H]\end{cases} \tag{3}$$

迭代策略只需建立一个模型，计算量小，然而用预测值作为模型输入，将导致预测误差的累积，从而影响预测精度。

2.1.2　直接策略

直接策略首先由 Cox 提出[6]，它通过历史观测值为每个步长建立一个预测模型，在输入和 H 个输出中训练 H 个不同的预测模型分别预测 $s_{t+h}(h=1,2,\cdots,H)$。

直接策略首先将一维原始时间序列数据组织成 H 个数据集

$$D_1=\left\{(\boldsymbol{x}_t,y_{t1})\in\left(\mathbf{R}^P\times\mathbf{R}\right)\right\}^N,\cdots,D_H=\left\{(\boldsymbol{x}_t,y_{th})\in\left(\mathbf{R}^P\times\mathbf{R}\right)\right\}^N \tag{4}$$

其中，$\boldsymbol{x}_t=\{s_t,\cdots,s_{t-P+1}\}^P$，$y_{th}=s_{t+h}$。

然后，直接预测策略在数据集 $D_h\in\{D_1,\cdots,D_H\}$ 上分别训练 H 个不同的模型，即

$$y_{th}=f_h(\boldsymbol{x}_t)+\omega_h,h\in\{1,\cdots,H\} \tag{5}$$

模型训练完成后，得到未来 H 个步长的预测值

$$\hat{s}_{t+h}=\hat{f}_h\left(s_t,s_{t-1},\cdots,s_{t-P+1}\right),h\in\{1,\cdots,H\} \tag{6}$$

在直接策略中，预测值不再作为模型的输入，所以预测误差不会累积到下一步的预测中，因此预测精度一般比迭代策略有所提升。但这种策略与迭代策略相比计算量稍大。

2.1.3　MIMO 策略

MIMO 策略首先由 Bontempi 提出[7]，是一种结构为多输入多输出的策略。在这种策略中，预测值

不再是一个值，而是由未来时刻的 H 个预测值组成的一个向量。MIMO 策略应用一个多输出模型得到 H 个未来时刻的预测值。

MIMO 策略首先将一维原始时间序列数据组织成数据集

$$D=\left\{\left(\boldsymbol{x}_t,\boldsymbol{y}_t\right)\in\left(\mathbf{R}^P\times\mathbf{R}^H\right)\right\}^N \tag{7}$$

其中，$\boldsymbol{x}_t=\left\{s_t,\cdots,s_{t-P+1}\right\}^P$，$\boldsymbol{y}_t=\left\{s_{t+1},\cdots,s_{t+H}\right\}^H$。

然后开始训练多输出预测模型

$$\boldsymbol{y}_t=f\left(\boldsymbol{x}_t\right)+\omega \tag{8}$$

模型训练完成后，得到接下来的 H 个数据的预测值

$$\left\{\hat{s}_{t+1},\cdots,\hat{s}_{t+H}\right\}=\hat{f}\left(s_t,\cdots,s_{t-P+1}\right) \tag{9}$$

MIMO 策略事实上仍然可以被分解为多个模型，所以其计算量等同于建立 H 个模型，计算代价仍然比较高。

预测策略的代表性研究工作包括文献[8~10]。目前，很少有论文将上述预测策略应用于传感网环境监测时间序列数据进行预测性能的评估，本文通过大量的实验来验证 3 种策略在环境监测数据上的预测性能。

2.2 预测模型

在预测模型方面，目前流行的建模技术有自回归（AR）模型[11]，支持向量机（SVM）[12,13]和神经网络（NN, neural network）[14,15]等。前 2 种模型不能同时应用上文中 3 种主要的预测策略进行多步预测，并且 AR 模型只适用于线性时间序列的建模，而环境监测时间序列数据通常是非线性的。对于 NN 模型，其灵活的结构可以采用所有 3 种预测策略[16]，因此广泛用于多步预测领域。例如，文献[17]应用 NN 模型对道路交通信息数据实现了多步预测；文献[18]对 NN 建模技术进行了多步预测的预测性能研究；在文献[19]中，作者将 NN 模型应用于油价的多步预测；文献[20]将多项式神经网络和迭代预测策略相结合对设备状态进行了多步预测；文献[21]提出先对时间序列进行 EMD 分解得到若干个分量，然后用神经网络模型分别对时间序列的 EMD 分量进行多步预测，最后将该若干个分量作为输入，用神经网络模型进行时间序列的多步预测。然而，神经网络的结构设计尚没有统一的理论依据，比如网络的层数、神经元的个数、传递函数的选择等，因此网络构建上有一定的困难。

文献[22]提出用高斯过程模型对混沌时间序列进行单步与多步预测，并提出用混合的高斯过程核函数进行建模。较之 NN 模型，高斯过程模型具有完备的理论基础，可以通过贝叶斯方程进行描述。尽管如此，文献[22]提出的方法（包括基于 NN 模型的方法），都是根据数据的统计特征建立模型，没有将数据具有的物理意义考虑到模型的设计当中，而传感网环境监测数据通常具有明确的物理意义，因此现有的相关工作对该类数据的建模不具有针对性。本文针对环境参数的物理特性，提出了高斯过程建模方法，为高斯过程模型构建了能够描述该类环境参数物理特性的核函数，从而提高了多步预测的性能。

3 高斯过程预测模型

3.1 高斯过程建模

将利用历史环境监测时间序列数据预测未来时刻的数据值的问题转化为回归问题，通过高斯过程建模，建立输入值（历史时间序列，长度为回归阶数 P）与输出值（H 步之后的未来时刻的数据值）之间的映射关系[22]。

给定环境监测时间序列 $s_1,\cdots,s_t$，构建 N 个训练输入数据 $\boldsymbol{X}=\left[\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N\right]$，$\boldsymbol{x}_i=\left\{s_i,\cdots,s_{i-P+1}\right\}$，$\boldsymbol{X}$ 存在于一个输入空间 $\boldsymbol{X}\in\mathbf{R}^P$。第 i 个训练输入 $\boldsymbol{x}_i$ 与一个训练输出 y_i 相对应，$y_i=s_{i+H}$，$y_i\in\mathbf{R}$，所有训练输出值构成一个向量 $\boldsymbol{y}=\left[y_1,\cdots,y_N\right]^{\mathrm{T}}$。

假设观测目标值 y_i 由一个未知函数 f 决定，并且被一个独立同分布的高斯噪声 ω_i 腐蚀[22]，即

$$y_i=f\left(\boldsymbol{x}_i\right)+\omega_i \tag{10}$$

其中，ω_i 为独立的随机变量，符合高斯分布，均值为 0，方差为 σ^2（假设训练和测试数据的噪声具有相同的分布），即 $\omega_i\sim N\left(0,\sigma^2\right)$。

在高斯过程回归方法中，通常赋予 f 一个均值为 0 的高斯过程先验概率分布[22]

$$f\left|\boldsymbol{x},\theta\sim N\left(0,\boldsymbol{K}\right)\right. \tag{11}$$

其中，$\boldsymbol{K}$ 是一个 $N\times N$ 维的对称正定的协方差矩阵，$K_{ij}=k\left(x_i,x_j\right)$，$k(\cdot,\cdot)$ 是协方差函数，也称核函数。它是一个以超参数 θ（在下文中进行介绍）为自变量的正定函数。因此，训练输出 y 的概率分布为

$$y\left|\boldsymbol{x},\theta\sim N\left(0,\boldsymbol{K}+\sigma^2\boldsymbol{I}\right)\right. \tag{12}$$

给定一个测试输入数据 $\boldsymbol{x}_*$，和对应的测试输出数据 y_* 以及相应的根据式(11)的先验概率分布抽样得到 $f(\boldsymbol{x}_*)$，$f(\boldsymbol{x})$ 和 $f(\boldsymbol{x}_*)$ 的联合概率分布也是一个均值为 0 的多维高斯过程，如式(13)所示[23]

$$\left[\begin{matrix}f(\boldsymbol{x})\\ f(\boldsymbol{x}_*)\end{matrix}\right]\Bigg|\boldsymbol{x},\boldsymbol{x}_*,\theta\sim N\left(0,\left[\begin{matrix}\boldsymbol{K} & \boldsymbol{K}_*\\ \boldsymbol{K}_*^{\mathrm{T}} & K_{**}\end{matrix}\right]\right) \tag{13}$$

其中，$\boldsymbol{K}_*=\left[k(\boldsymbol{x}_*,\boldsymbol{x}_1),\cdots,k(\boldsymbol{x}_*,\boldsymbol{x}_N)\right]^{\mathrm{T}}$，$K_{**}=k(\boldsymbol{x}_*,\boldsymbol{x}_*)$。

基于式(10)中的高斯噪声假设，训练输出 y 和测试输出 y_* 的联合概率分布如式(14)所示[23]

$$\left[\begin{matrix}y\\ y_*\end{matrix}\right]\Bigg|\boldsymbol{x},\boldsymbol{x}_*,\theta\sim\left(0,N\left[\begin{matrix}\boldsymbol{K}+\sigma^2\boldsymbol{I} & \boldsymbol{K}_*\\ \boldsymbol{K}_*^{\mathrm{T}} & K_{**}+\sigma^2\end{matrix}\right]\right) \tag{14}$$

根据条件高斯法则，由式(14)可以推导出[23]

$$y_*\left|y,\boldsymbol{x},\theta,\sigma^2\sim N\left(m(\boldsymbol{x}_*),\upsilon(\boldsymbol{x}_*)\right)\right. \tag{15}$$

其中，预测值 y_* 的均值 $m(\boldsymbol{x}_*)$ 和标准差 $\upsilon(\boldsymbol{x}_*)$ 为[23]

$$m(\boldsymbol{x}_*)=\boldsymbol{K}_*^{\mathrm{T}}\left(\boldsymbol{K}+\sigma^2\boldsymbol{I}\right)^{-1}y \tag{16}$$

$$\upsilon(\boldsymbol{x}_*)=K_{**}+\sigma^2-\boldsymbol{K}_*^{\mathrm{T}}\left(\boldsymbol{K}+\sigma^2\boldsymbol{I}\right)^{-1}\boldsymbol{K}_* \tag{17}$$

由此可以看出，已知协方差函数和它的超参数，利用式(16)就可以对未来时刻的数据做出预测。

3.2　GP 核函数的构建

GP 可选择不同的协方差函数（也称核函数），每一种协方差函数都对应若干个超参数。对于某个具体的应用，确定了协方差函数的函数形式以及函数对应的所有超参数的值也就得到了高斯过程模型的具体形式[24]。

2 个协方差函数进行相加或相乘操作后仍然是一个协方差函数[23,25]。因此，可以将几个基础协方差函数通过相加和（或）相乘的操作创建一个新的协方差函数。例如，任意一个核函数 S 都可以用 $S+B$ 或者 $S\times B$ 来代替，并且任意一个基础核函数 B 都可以被任意一个其他的基础核函数 B' 来替换。

基于以上准则，基于以下 4 种常用的核函数来构建针对环境监测时间序列数据的高斯过程核函数，包括：平方指数核函数（SE）、周期核函数（PER）、线性核函数（LIN）、二次有理数核函数（RQ）。其具体的函数形式和相应的超参数[26]如下

$$k_{\mathrm{SE}}(x,\hat{x})=\sigma^2\exp\left(-\frac{(x-\hat{x})^2}{2l^2}\right) \tag{18}$$

平方指数核函数 SE 的超参数 $\theta_{\mathrm{SE}}=(\sigma,l)$。

$$k_{\mathrm{PER}}(x,\hat{x})=\sigma^2\exp\left(-\frac{2\sin^2\left(\pi|x-\hat{x}|/p\right)}{l^2}\right) \tag{19}$$

周期核函数 PER 的超参数 $\theta_{\mathrm{PER}}=(\sigma,p,l)$。

$$k_{\mathrm{LIN}}(x,\hat{x})=\sigma_c^2+\sigma_v^2(x-s)(\hat{x}-s) \tag{20}$$

线性核函数 LIN 的超参数 $\theta_{\mathrm{LIN}}=(\sigma_c,\sigma_v,s)$。

$$k_{\mathrm{RQ}}(x,\hat{x})=\sigma^2\left(1+\frac{(x-\hat{x})^2}{2\alpha l^2}\right)^{-\alpha} \tag{21}$$

二次有理数核函数 RQ 的超参数 $\theta_{\mathrm{RQ}}=(\sigma,\alpha,l)$。

其中，SE 描述数据局部变化的特征，PER 描述数据周期性变化的特征，并且是一个全局性的周期核函数，LIN 描述数据的长期变化趋势，RQ 描述数据的不规则变动。

基于上述准则，提出一种新的针对环境监测数据特性的核函数构建方法。

给定环境监测历史数据，首先，通过经验模态分解（EMD），获得数据的多个 IMF（intrinsic mode functions）分量，分析其数据特性（如周期性）。接着，根据需要建模的数据拥有的特性，通过相加或者相乘操作融入对应的基本核函数，使最终的核函数能够表述数据特性。具体操作包括：当需要建模的数据具有周期性时，则通过相加操作在最终的核函数中融入 PER 核函数；当数据的某个特性是局部的而非全局的特性时，则通过与 SE 核函数的相乘操作，把一个全局的结构特性转变为局部的结构特性[26]；当需要建模的数据不规则变动时，则通过相加操作在最终的核函数中融入 RQ 核函数。请注意，上述过程是对历史数据的特性进行建模，是对数据先验知识的描述，而预测对象是未来时刻的数据值，是对数据后验知识的估计。将在第 4 节给出具体的核函数构建的例子。

3.3　超参数的求解

本节通过极大似然法获得超参数的最优取值，即通过建立训练样本条件概率的对数似然函数，并对超参数求偏导,再采用共轭梯度优化方法搜索出超参数的最优解，具体求解过程可见参考文献[23]。其中对数似然函数的形式为

$$\log p\left(\boldsymbol{y}\middle|\boldsymbol{x},\theta\right)=-\frac{1}{2}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}_y^{-1}\boldsymbol{y}-\frac{1}{2}\log\left|\boldsymbol{K}_y\right|-\frac{N}{2}\log 2\pi \quad (22)$$

对上式求偏导的结果为

$$\frac{\partial}{\partial\theta_j}\log\left(\boldsymbol{y}\middle|\boldsymbol{x},\theta\right)=\frac{1}{2}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}_y^{-1}\frac{\partial\boldsymbol{K}_y}{\partial\theta_j}\boldsymbol{K}_y^{-1}\boldsymbol{y}-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{K}_y^{-1}\frac{\partial\boldsymbol{K}_y}{\partial\theta_j}\right)=\frac{1}{2}\mathrm{tr}\left(\left(\boldsymbol{\alpha}\boldsymbol{\alpha}^{\mathrm{T}}-\boldsymbol{K}_y^{-1}\right)\frac{\partial\boldsymbol{K}_y}{\partial\theta_j}\right) \quad (23)$$

其中，$\boldsymbol{K}_y=\boldsymbol{K}+\sigma^2\boldsymbol{I},\boldsymbol{\alpha}=\boldsymbol{K}_y^{-1}\boldsymbol{y}$。

4 实验验证

4.1 对比方法

本节将提出的高斯过程模型与 BP 神经网络和小波神经网络模型进行对比，并与 2.1 节提出的 3 种预测策略相结合，共形成 7 种多步预测方法：高斯过程迭代策略模型（GPI）、高斯过程直接策略模型（GPD）、BP 迭代策略模型（BPI）、BP 直接策略模型（BPD）、BP 多维输入输出模型（BPM）、小波迭代策略模型（WNI）和小波直接策略模型（WND）。

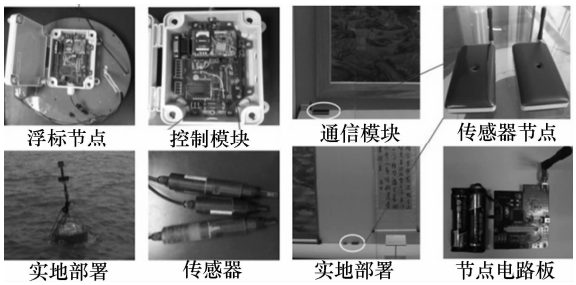
4.2 实验参数设置

对于每种实验数据，根据建模方法和预测策略，将其组织成输入输出数据对，形成原始数据集，并将其随机划分为训练数据集和测试数据集。基于训练数据集训练预测模型，并将训练得到的模型在测试数据集上进行性能测试。上述过程重复 20 次，求取预测性能的平均结果。

高斯过程模型的核函数选择见 4.5 节。在应用 BP 神经网络模型时，根据文献[18]，神经网络层数的设置没有通用的准则，一个 3 层的神经网络结构就能够描述数据的线性和非线性关系，根据文章提到的 Kolmogorovs 理论，若输入层的节点个数为 n，则当隐藏层节点个数大于等于 $2n+1$ 时可满足输入层的任一函数，本文同样应用该理论，实验中 n 为 4，当设置隐藏层的神经元个数为 10 时可满足要求，同时设置隐藏层的传递函数为 tansig 函数，输出层的传递函数为 purelin 函数，BP 网络的训练函数为 trainlm 函数。在应用小波神经网络模型时选择的小波基函数为 Morlet 小波 $\cos(rt)\mathrm{e}^{-\frac{t^2}{2}}$。

4.3 实验数据

实验所用数据来自 3 个数据集，其中 2 种数据集采集自实地部署的传感网环境监测应用系统，最后一种数据集来源于伯克利实验室在公开网站上公布的环境监测数据[27]。



(a) DS_1 部署 (b) DS_2 部署

图 1 DS_1 和 DS_2 的系统部署

数据集 1(DS_1)：图 1 (a)是部署的一个湖泊水质监测系统，用以采集湖体的水质参数，采集周期为 10 min。实验中用到的环境监测数据有：水温（记为 $DS_{1\text{-WT}}$）、PH（记为 $DS_{1\text{-PH}}$）和溶解氧（记为 $DS_{1\text{-OXY}}$），每种实验数据选取约 5 000 个数据点。

数据集 2(DS_2)：图 1(b)是部署的一个博物馆文物监测系统，用于采集展陈室的环境数据，采集周期为 15 min。实验中用到的环境监测数据有：温度（记为 $DS_{2\text{-T}}$）和湿度（记为 $DS_{2\text{-H}}$）等，每种实验数据选取约 5 000 个数据点。

数据集 3(DS_3)：该数据由伯克利实验室部署的试验床采集得到，系统由 54 个 Mica2 传感器节点组成。在实验中选用了节点 10 的 3 000 个温度数据（记为 $DS_{3\text{-T}}$）和 3 000 个湿度数据（记为 $DS_{3\text{-H}}$），采数间隔为 31 s。

4.4 评价指标

本节使用均方根误差、累积相对预测误差来评价多步预测方法的性能。

均方根误差（RMSE, root mean square error）又叫标准误差，其计算表达式如下

$$RMSE=\sqrt{\frac{\sum_{i=1}^{m}\left(y_i-\hat{y}_i\right)^2}{m}} \quad (24)$$

其中，y_i 表示实际值，$\widehat{y_i}$ 表示预测值，m 表示预测样本数。

累积相对预测误差百分比（CRPEP, cumulative relative prediction error percent)是指预测相对误差小于某个阈值的样本的数量占预测样本总数的百分比。其计算表达式如下

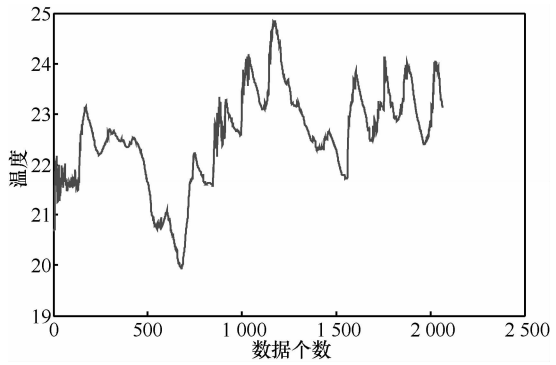
$$CRPEP = \frac{COUNT\left(\dfrac{|y_i - \hat{y}_i|}{y_i} < \delta_{\mathrm{RE}}\right)}{m} \tag{25}$$

其中，y_i、$\hat{y}_i$、m 与式(24)同义，δ_{RE} 为误差阈值。$COUNT(F)$ 函数返回满足条件 F 的样本数量。

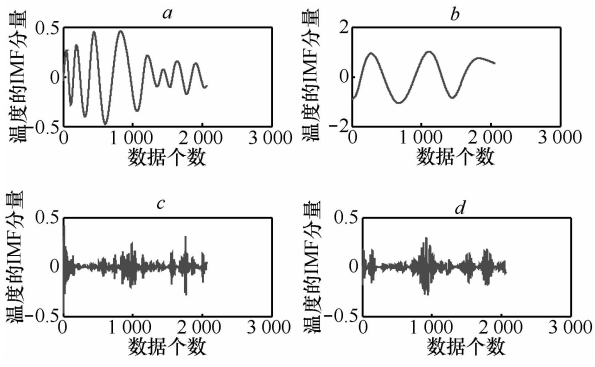
4.5 GP 核函数的构建

在传感网环境监测系统中，传感器采集的环境监测数据具有一些数据特性。例如数据在时间上呈现一定的周期性，但并不是规则的周期变化，在一小段时间内存在着局部的变化和不规则的跳跃等。介于篇幅，以环境参数 $DS_{1\text{-WT}}$ 数据为例，介绍针对环境监测数据物理特性的核函数的构建方法。

图 2 (a)是 $DS_{1\text{-WT}}$ 的原始数据曲线 T，通过对 T 做经验模态分解，可以得到 T 的若干个 IMF（intrinsic mode functions）分量[18,28]，如图 2(b)所示。由图 2 可以看出，首先，环境参数存在周期性的变化趋势（如图 2(b)的 a 分量和 b 分量所示），因此，通过相加操作在最终的核函数中融入 PER 核函数；其次，图 2(a)的原始曲线是局部周期性的曲线，所以通过 PER 与 SE 核函数的相乘操作，在最终的核函数中引入 $PER \times SE$ 的形式；再次，图 2(b)的 c 分量和 d 分量表现了环境参数的局部变化和不规则变化的特性，通过加入 SE 核函数来表示数据的局部变化特性，通过加入 RQ 核函数来表示数据的不规则变化；最后，由于环境数据中通常存在着噪声信号，通过加入 SE 核函数（记为 SE_{noise}）来描述噪声的局部变化。综上所述，构建的核函数的最终形式为



(a) $DS_{1\text{-WT}}$ 的数据曲线



(b) $DS_{1\text{-WT}}$ 的部分 IMF 分量

图 2　$DS_{1\text{-WT}}$ 的原始数据及 IMF 分量

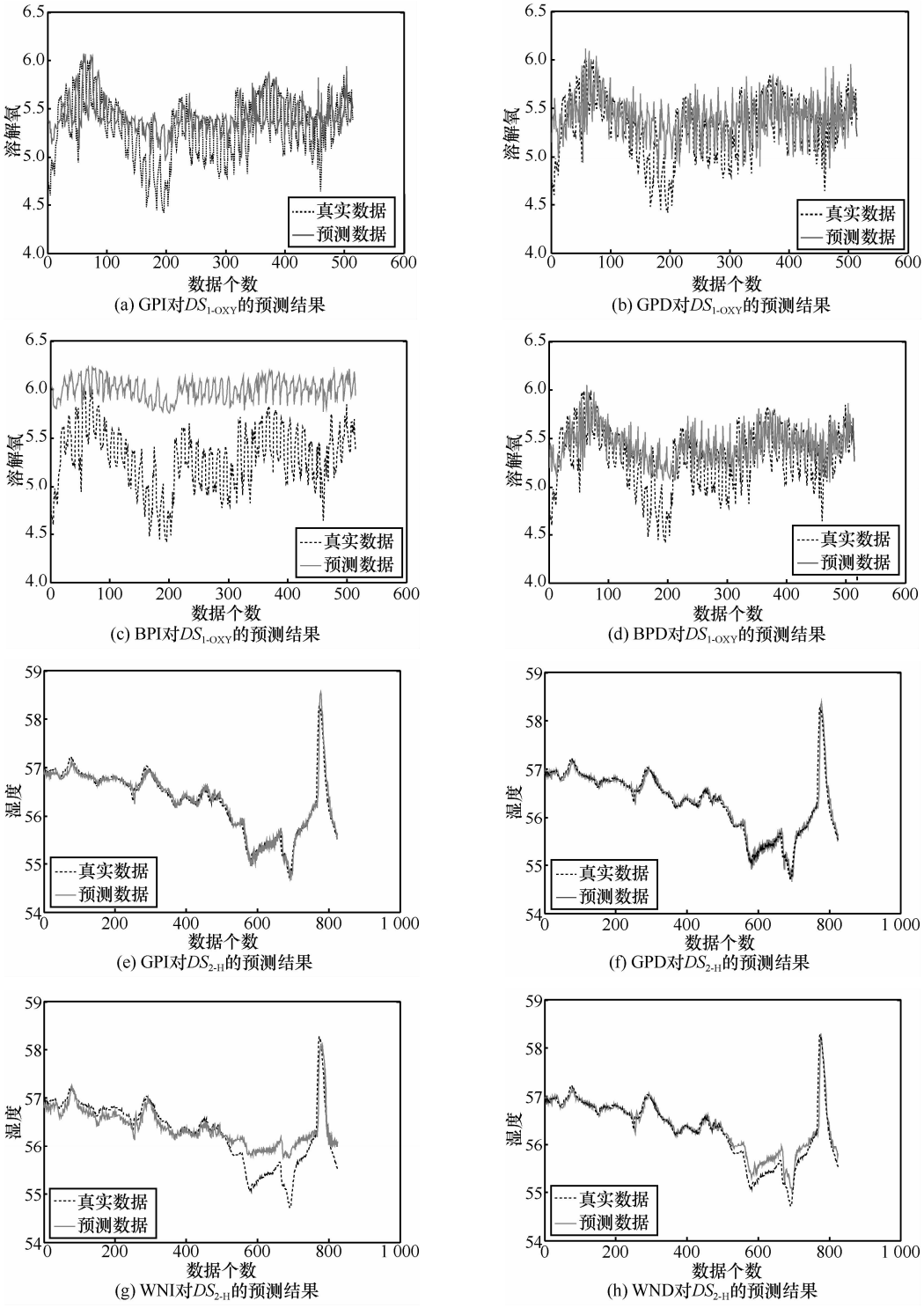
$$K = SE + PER \times SE + RQ + SE_{\mathrm{noise}} \tag{26}$$

其他环境数据的核函数构建过程类似，不再赘述。

4.6 实验结果分析

在这节中，首先，保持回归阶数 P 和多步预测步长 H 不变，在 3 种数据集上测试 7 种预测方法的性能。然后，通过调整预测步长 H 来测试预测步长对几种预测方法的预测性能的影响。最后，在回归阶数 P 和多步预测步长 H 不变的情况下，研究数据采样间隔对几种预测方法性能的影响。由于篇幅有限，只列举了一部分数据的实验结果，其他结果具有类似的预测精度。

基于数据 $DS_{1\text{-OXY}}$ 和数据 $DS_{2\text{-H}}$，给出了几种预测方法得到的预测数据与真实数据的对比曲线，如图 3 所示，其中，回归阶数 $P=4$，预测步长 $H=5$。从图 3 中可以看出，与其他方法相比，基于高斯过程模型的预测方法得到的预测值曲线与真实值曲线拟合得更好，能够更好地适应数据的变化，而其他方法对数据变化的反应较慢，并且预测结果滞后于真实值。尤其在数据时常发生剧烈变化的情况下，如数据 $DS_{1\text{-OXY}}$，基于高斯过程模型的预测方法无论是与迭代策略还是与直接策略相结合，其预测结果都明显优于基于 BP 神经网络的方法，从而说明提出的基于高斯过程的预测方法具有更高的预测精度。从预测策略上分析，小波神经网络迭代策略模型的预测数值偏离了真实值，而小波神经网络直接策略模型由于没有预测误差的累积，其预测性能要比迭代策略好。所以，从总体上看，针对同一模型而言，直接策略的预测精度要高于迭代策略的精度。在预测模型方面，由于提出的高斯过程模型通过分析数据的物理特性，构建了描述相应物理特性的核函数，因此，当训练模型参数时，可以同时根据数据统计特性和物理特性来训练，而神经网络仅仅基于数据的统计特性训练模型，因此



(a) GPI对$DS_{1\text{-}OXY}$的预测结果　(b) GPD对$DS_{1\text{-}OXY}$的预测结果

(c) BPI对$DS_{1\text{-}OXY}$的预测结果　(d) BPD对$DS_{1\text{-}OXY}$的预测结果

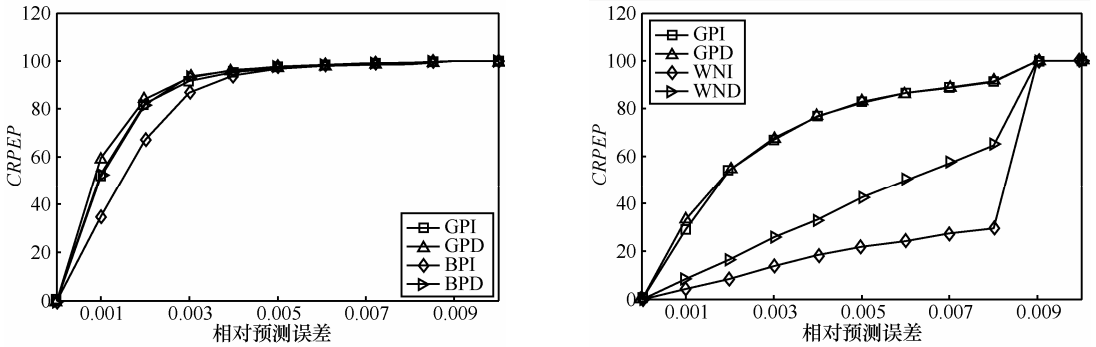
(e) GPI对$DS_{2\text{-}H}$的预测结果　(f) GPD对$DS_{2\text{-}H}$的预测结果

(g) WNI对$DS_{2\text{-}H}$的预测结果　(h) WND对$DS_{2\text{-}H}$的预测结果

图 3　4 种预测方法的对比曲线

其预测性能低于提出的高斯模型。

图 4 表示的是 4 种预测方法在数据 $DS_{2\text{-}H}$ 和数据 $DS_{3\text{-}T}$ 上的累积相对预测误差百分比（CRPEP）的分布情况。其中，回归阶数 $P=4$，预测步长 $H=5$。由图 4(a)可以看出，GPD 预测方法 60%的预测数据的相对误差在 0.001 以下，而 BPD 预测方法和 GPI 方法的预测性能类似，有不到 55%的预测数据的相对误差小于 0.001，BPI 方法由于误差的传播导致其相对误差在 0.001 以下的样本只有 38%。在图 4(b)中，GPI 与 GPD 方法的预测精度类似，WND 的预测精度要优于 WNI 的预测精度，基于 GP 的预测方法在预测精度上显著优于基于 WN 的



(a) 在$DS_{2\text{-H}}$数据集上预测结果的*CRPEP*　　(b) 在$DS_{3\text{-T}}$数据集上预测结果的*CRPEP*

图 4　4 种预测方法的 *CRPEP* 曲线

预测方法。

7 种预测方法对 3 种数据集的预测性能（通过 *RMSE* 计算得到）的统计结果如表 1 和表 2 所示，其中回归阶数 $P=4$，预测步长 $H=5$。由表 1 和表 2 可以看出，基于高斯过程的预测方法在绝大多数的数据集上都有最小的 *RMSE*。这说明提出的基于高斯过程的预测方法得到的预测值与其他方法相比更接近于真实的环境监测数据，其预测精度比对比方法提高了 20%（通过计算 *RMSE* 的平均降低率得到，例如对于数据 $DS_{1\text{-WT}}$，方法 GPI 与方法 BPI 相比，*RMSE* 降低率为 $\frac{0.214\,9-0.170\,6}{0.214\,9}\times 100\%=20\%$）。针对同一模型的不同策略的预测结果可以看出，在相同数据的多步预测上，由于直接预测策略不存在误差累积问题，其预测精度要高于相同模型的直接策略。从 BP 神经网络模型与 3 种预测策略结合的预测方法分析，可以看出，BPD 与 BPM 策略在不同的数据集上有着不同的预测性能，但总体上比 BPI 方法预测精度高。

表 1　迭代策略在不同数据集上预测结果的 *RMSE*

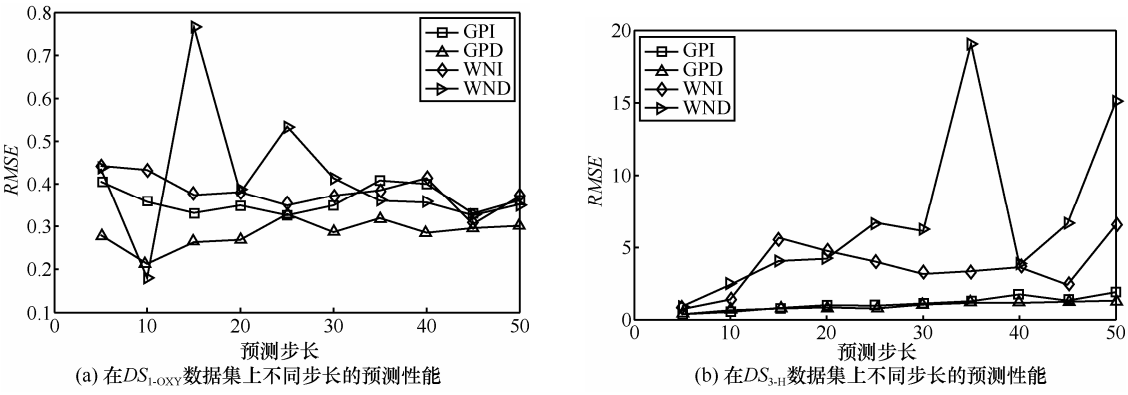
数据	方法		
	GPI	BPI	WNI
$DS_{3\text{-H}}$	0.393 5	0.433 3	1.805 2
$DS_{3\text{-T}}$	0.179 3	0.203 5	1.064 0
$DS_{1\text{-OXY}}$	0.350 3	0.392 4	0.466 5
$DS_{1\text{-PH}}$	0.030 0	0.032 2	0.044 7
$DS_{1\text{-WT}}$	0.170 6	0.214 9	0.305 9
$DS_{2\text{-H}}$	0.333 6	0.353 8	0.395 5
$DS_{2\text{-T}}$	0.162 8	3.972 7	1.129 2

表 2　直接策略在不同数据集上预测结果的 *RMSE*

数据	方法			
	GPD	BPD	WND	BPM
$DS_{3\text{-H}}$	0.386 1	0.388 9	0.461 9	0.407 0
$DS_{3\text{-T}}$	0.179 2	0.187 4	0.626 6	0.192 9
$DS_{1\text{-OXY}}$	0.313 4	0.320 0	0.377 6	0.322 6
$DS_{1\text{-PH}}$	0.029 8	0.034 2	0.035 0	0.033 3
$DS_{1\text{-WT}}$	0.180 3	0.201 5	0.222 3	0.178 1
$DS_{2\text{-H}}$	0.288 3	0.313 8	0.327 9	0.415 4
$DS_{2\text{-T}}$	0.153 5	0.184 7	0.211 39	0.214 5

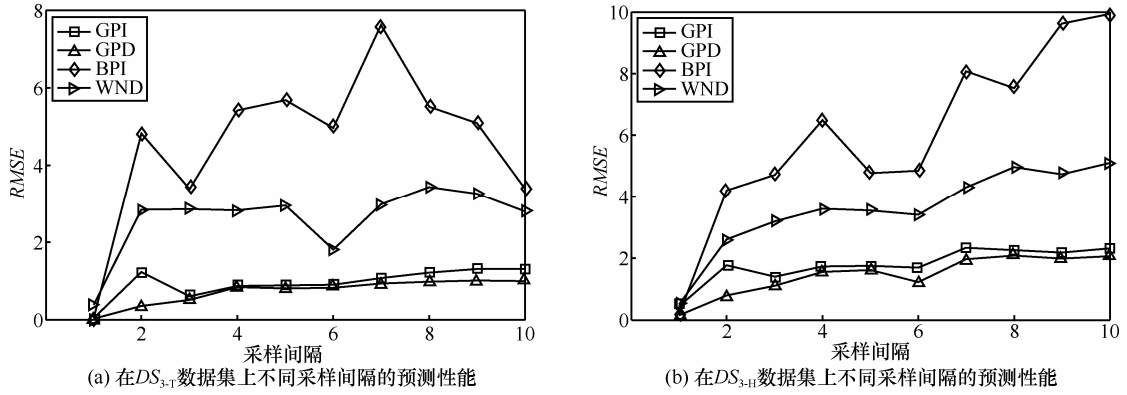
其次，基于数据 $DS_{1\text{-OXY}}$ 和数据 $DS_{3\text{-H}}$，研究预测步长 H 对几种预测方法的预测性能的影响。固定回归阶数 $P=4$，预测步长从 5 变化到 50，递增幅度为 5，相应的预测结果的 *RMSE* 的变化情况如图 5 所示。可以看出，随着预测步长的增大，各种方法的预测结果的 *RMSE* 呈逐渐增大的趋势。特别地，GPD 预测方法的 *RMSE* 曲线始终位于其他预测方法的 *RMSE* 曲线之下，表明在各种预测步长之下，提出的基于高斯过程的直接预测方法能够对环境监测数据进行更准确的多步预测。其平均预测精度较之 WN，可以提高约 40%（通过计算 *RMSE* 的平均降低率得到，例如对于数据 $DS_{3\text{-H}}$，步长为 5 时，GP 与 WN 相比，*RMSE* 降低率为 $\frac{0.897\,3-0.363\,8}{0.897\,3}\times 100\%=59\%$）。

最后，研究数据采样间隔对预测性能的影响。通过对数据集 $DS_{3\text{-T}}$ 和数据集 $DS_{3\text{-H}}$ 的原始数据进行抽样，达到改变数据采样间隔的目的。具体来说，给定原始数据的采样间隔 W，通过改变抽样频率，得到了采样间隔从 W 到 $10W$、递增幅度为 W 的数据集合。固定回归阶数 $P=4$，预测步长 $H=10$，相应预测结果的 *RMSE* 随数据采样间隔的变化情况如图 6 所示。整体上看，随着采样间隔的增大，各种方法的预测结果的 *RMSE* 呈逐渐



(a) 在$DS_{1\text{-}OXY}$数据集上不同步长的预测性能　　(b) 在$DS_{3\text{-}H}$数据集上不同步长的预测性能

图 5　预测方法在不同预测步长上的性能对比



(a) 在$DS_{3\text{-}T}$数据集上不同采样间隔的预测性能　　(b) 在$DS_{3\text{-}H}$数据集上不同采样间隔的预测性能

图 6　预测方法在不同采样间隔上的性能对比

增大的趋势。这是因为随着采样间隔的增大，数据提供的特征信息逐渐减少，从而降低了模型的预测精度。然而，在预测精度的稳定性上，随着采样间隔的增大，BP 神经网络和小波神经网络预测结果的 *RMSE* 变化较大，而 GP 预测结果的 *RMSE* 变化幅度较低且变化稳定。此外，GP 方法预测结果的 *RMSE* 曲线始终位于其他方法的 *RMSE* 曲线之下，即具有更高的预测精度。由此可知，与其他方法相比，GP 预测方法对不同的数据采样间隔具有更强的适应性。

综上所述，与几种对比方法相比，基于高斯过程的多步预测方法具有更高的预测精度。这主要是因为高斯过程建模可以通过核函数描述建模数据的特性，针对环境监测数据的物理特性构建的核函数能够更好地描述环境监测数据的内在变化规律，而神经网络模型未能在建模过程中引入针对性的数据特性表示方法，因此在预测性能上有所下降。

4.7　高斯过程不同核函数预测性能比较

在这一节中，使用 3 种不同的核函数分别建立高斯过程预测模型，包括：1）应用最广泛的平方指数核函数；2）文献[22]中的核函数（式（19））；3）根据 3.2 节提出的核函数构建方法得到的核函数。根据他们的预测性能，验证了提出的核函数构建方法的有效性。

将基于上述 3 种核函数的高斯过程模型记为 GPN、GPL、GP，分别与迭代策略和直接策略相结合，得到 6 种多步预测方法，记为 GPNI、GPND、GPLI、GPLD、GPI、GPD。

图 7 给出了基于 GPN 和 GP 模型的 4 种方法对 $DS_{3\text{-}T}$ 数据的预测值与真实值的对比曲线。表 3 和表 4 给出了上述 6 种预测方法在 5 个数据集上的预测性能（通过 *RMSE* 计算得到）。其中，回归阶数 $P=4$，预测步长 $H=10$。

结合图 7、表 3 和表 4 可以看出，基于得到的核函数所建立的高斯过程预测方法，在绝大多数数据集上都具有最小的 *RMSE*，预测值曲线与真实值曲线的拟合度更好。其原因如下：普通高斯过程模型仅仅基于数据的统计特性训练模型参数，而提出的高斯过程模型通过分析时间序列数据的物理特性，构建描述相应物理特性的核函数，因此可以同时根据数据统计特性和物理特性来训练模型参数。综上所述，提出的核函数构建方法可以描述

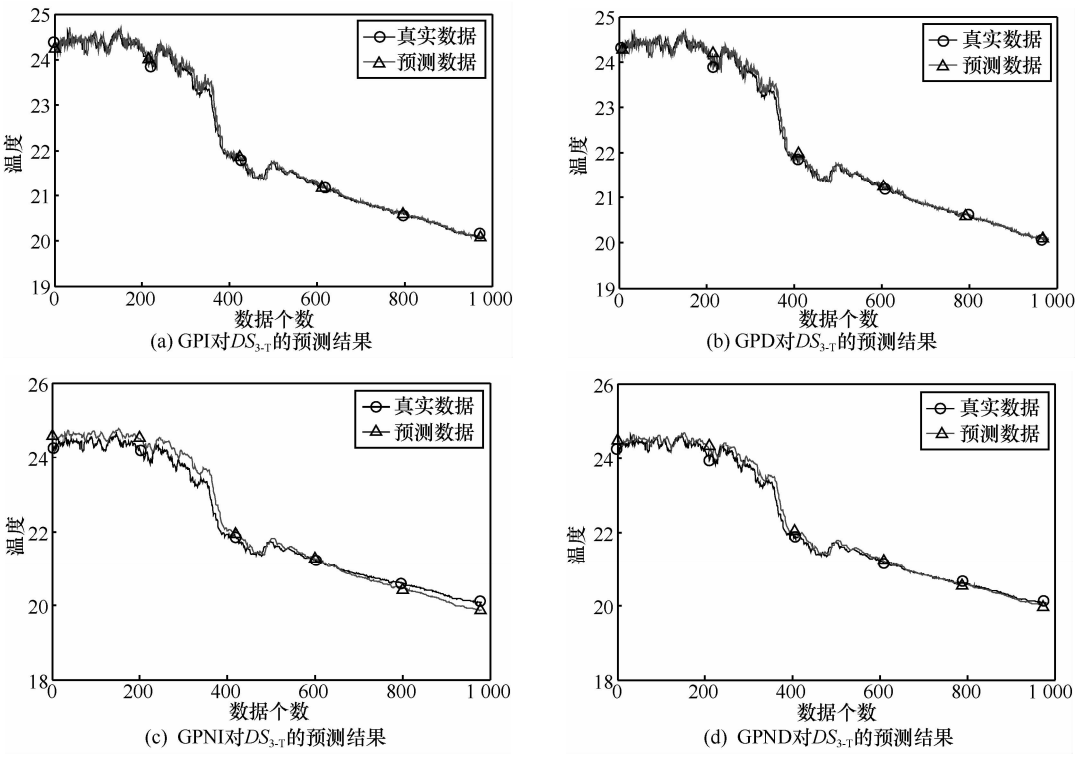


图 7　4 种预测方法的对比曲线

数据特性，获得更好的预测结果。

表 3　GPN 与 GP 在数据集上预测结果的 *RMSE*

数据	方法		
	GPNI	GPND	GPI
$DS_{3\text{-H}}$	1.157 8	0.592 0	0.393 5
$DS_{1\text{-OXY}}$	0.532 2	0.472 8	0.350 3
$DS_{1\text{-PH}}$	0.051 5	0.049 9	0.030 0
$DS_{1\text{-WT}}$	0.213 6	0.203 2	0.170 6
$DS_{2\text{-T}}$	0.368 9	0.286 7	0.162 8

表 4　GPL 与 GP 在数据集上预测结果的 *RMSE*

数据	方法		
	GPD	GPLI	GPLD
$DS_{3\text{-H}}$	0.386 1	0.506 2	0.502 0
$DS_{1\text{-OXY}}$	0.323 4	0.534 4	0.472 8
$DS_{1\text{-PH}}$	0.029 8	0.048 4	0.050 0
$DS_{1\text{-WT}}$	0.180 3	0.201 0	0.205 1
$DS_{2\text{-T}}$	0.153 5	0.265 7	0.223 7

5　结束语

针对传感网环境监测系统采集的大量的时间序列数据，本文提出了基于高斯过程的多步预测方法，该方法通过对环境数据的 EMD 分解得到的 IMF 分量和数据本身具有的物理意义的分析，构建了适用于环境监测数据的高斯过程核函数来描述数据的特性；同时结合不同的多步预测策略，本文对 5 种环境监测数据进行了大量的实验以对比不同多步预测方法的预测性能。实验结果表明，本文提出的高斯过程多步预测方法能对环境数据变化实现准确的描述和预测，其预测精度比对比方法提高了 20%。

本文提出的基于高斯过程的多步预测方法具有通用性。具体来说，仅需要适当地重新构建核函数，无需改变模型的建模过程和超参数求解，即可应用于其他类型的时间序列数据的多步预测。将在进一步工作中验证其在多种时间序列数据（如溶解氧时间序列数据、GPS 时间序列数据等）多步预测中的性能。

参考文献：

[1] 崔莉, 鞠海玲, 苗勇, 等. 无线传感器网络研究进展[J]. 计算机研究与发展, 2005, 42(1):163-174.
CUI L, JU H L, MIAO Y, *et al*. Research overview of wireless sensor network[J]. Journal of Computer Research and Development, 2005, 42(1):163-174.

[2] 孙利民, 李建中, 陈渝, 等. 无线传感器网络[M]. 北京: 清华大学出版社, 2005.
SUN L M, LI J Z, CHEN Y, *et al*. Wireless Sensor Network[M]. Beijing: Press of Tsinghua University, 2005.

[3] BROCKWELLL P J, DAVIS R A. Introduction to Time Series and Forecasting[M]. New York: Springer, 1994.

[4] JIN L, JU J M, MIAO Q L. Study on Ann-based multi-step prediction

model of short-term climatic variation[J]. Advances Atmosphere Sciences, 2000, 17(1):157-164.

[5] CHEVILLON G. Direct muti-step estimation and forecasting[J]. Journal of Economic Surveys, 2007, 21(4): 746-785.

[6] COX D R. Prediction by exponentially weighted moving averages and related methods[J]. Journal of the Royal Statistical Society Seri B, 1961, 23(1): 414-442.

[7] BONTEMPI G. Long term time series prediction with multi-input multi-output local learning[A]. Proc of the 2nd European Symposium on Time Series Prediction[C]. Helsinki, Finland, 2008. 145-154.

[8] MCELROY T, WILDI M. Multi-step-ahead estimation of time series models[J]. International Journal of Forecasting, 2013, 29(3): 378-394.

[9] TAIEB S B, SORJAMAA A, BONTEMPI G. Multiple-output modeling for multi-step-ahead time series forecasting[J]. Neurocomputing, 2010, 73(10): 1950-1957.

[10] BAO Y, XIONG T, HU Z Y. Multi-step-ahead time series prediction using multiple-output support vector regression[J]. Neurocomputing, 2013, 129(10): 482-493.

[11] PROIETTI T. Direct and iterated multistep AR methods for difference stationary processes[J]. International Journal of Forecasting, 2011, 27(2): 266-280.

[12] HUANG Z F, SHYU M L. K-NN based LS-SVM framework for long-term time series prediction[A]. Proc of the IEEE International Conference on Information Reuse and Integration[C]. Las Vegas NV, USA, 2010. 69-74.

[13] AIGUO S A, BO Z. *K*-nearest neighbor LS-SVM method for multi-step prediction of chaotic time series[A]. 2012 IEEE Symposium on Electrical and Electronics Engineering[C]. Kuala Lumpur, Malaysia, 2012. 407-409.

[14] CHEN P A, CHANG L C, CHANG F J. Reinforced recurrent neural networks for multi-Step-ahead flood forecasts[J]. Journal of Hydrology, 2013, 497(8): 71-79.

[15] SHEN H Y, CHANG L C. Online multistep-ahead inundation depth forecasts by recurrent NARX networks[J]. Hydrology and Earth System Sciences, 2013, 17(3): 935-945.

[16] PILKA F, ORAVEC M. Multi-step ahead prediction using neural networks[A]. Proceedings of the 53rd International Symposium ELMAR[C]. Zadar, Croatia, 2011. 269-272.

[17] LIANG L J, YU D, CHANG X. Research on multi-step prediction of the short-term information by empirical model decomposition to abnormal state road traffic information[A]. Proc of 2012 International Conference on Computer Science and Service System[C]. Nanjing, China, 2012. 2034-2037.

[18] GUO Z H, ZHAO H Y, LU H Y, *et al*. Multi-step forecasting for wind speed using a modified EMD-based artificial neural network model[J]. Renew Energy, 2012, 37(1): 241-249.

[19] XIONG T, BAO Y K, HU Z Y. Beyond one-step-ahead forecasting: evaluation of alternative multi-step-ahead forecasting models for crude oil prices[J]. Energy Economics, 2013, 40(6): 405-415.

[20] 王秋香, 于德介. 设备状态的多项式神经网络迭代多步预测法[J]. 计算机仿真, 2010, 27(3): 179-181, 262.
WANG Q X, YU D J. Iterative multi-step based PFANN for condition prediction of equipment[J]. Computer Simulation, 2010, 27(3): 179-181, 262.

[21] 谢景新, 程春田, 周桂红, 等. 基于经验模式分解与混沌分析的直接多步预测模型[J]. 自动化学报, 2008, 34(6): 684-689.
XIE J X, CHENG C T, ZHOU G H, *et al*. A new direct multi-step ahead prediction model based on EMD and chaos analysis[J]. Acta Automatica Sinica, 2008, 34(6): 684-689.

[22] 李军, 张友鹏. 基于高斯过程的混沌时间序列单步与多步预测[J]. 物理学报, 2011, 60(7): 143-152.
LI J, ZHANG Y P. Single-step and multiple-step prediction of chaotic time series using Gaussian process model[J]. Acta Physica Sinica, 2011, 60(7): 143-152.

[23] RASMUSSEN C E, WILLIAMS C K I. Gaussian Processes for Machine Learning[M]. London: The MIT Press, 2006.

[24] OSBORNE M A, ROBERTS S J, ROGERS A, *et al*. Towords real-time information processing of sensor network data using computationally efficient multi-output gaussian process[A]. Proc of the International Conference on Information Processing in Sensor Networks[C]. St. Louis MO, USA, 2008. 109-120.

[25] WILLIAMS, C K I, BARBER, D. Bayesian classification with Gaussian processes[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(12):1342-1351.

[26] DUVENAUD D, LLOYD J R, GROSSE R, *et al*. Structure discovery in nonparametric regression through compositional kernel search[A]. Proc of the 30th International Conference on Machine Learning[C]. Atlanta GA, USA, 2013. 1166-1174.

[27] Intel Berkeley Lab[EB/OL]: http://db.lcs.mit.edu/labdata/labdata.html.

[28] MANDIC D, REHMAN N, WU Z, *et al*. Empirical mode decomposition-based time-frequency analysis of multivariate signals[J]. The Power of Adaptive Data Analysis, 2013, 30(6): 74-86.

作者简介：

陈艳（1990-），女，山东临沂人，中国科学院硕士生，主要研究方向为信息融合。

王子健（1980-），男，河北唐山人，中国科学院助理研究员，主要研究方向为无线传感器网络多元数据融合与智能处理。

赵泽（1978-），男，锡伯族，辽宁大连人，中国科学院高级工程师，主要研究方向为无线传感器网络和嵌入式系统。

李栋（1979-），男，黑龙江哈尔滨人，中国科学院副研究员，主要研究方向为物联网和传感器网络组网技术、物联网系统结构。

崔莉（1962-），女，北京人，中国科学院研究员，主要研究方向为传感器技术及无线传感器网络。