一种应用于时序随机潮流的鲁棒数据驱动过程建模

Pooja Algikar, *IEEE* 会员，Yijun Xu，*IEEE* 高级会员，Somayeh Yarahmadi，*IEEE* 会员，Lamine Mili，*IEEE* 终身会士。

arXiv:2301.02651v1 [eess.SY] 6 Jan 2023

摘要—**本文提出一种鲁棒数据驱动过程模型，其超参数采用 Schweppe 型广义最大似然估计量进行鲁棒估计。该模型基于记录的电压相量和功率注入时间序列数据进行训练，以执行时序随机潮流计算。电力系统数据常会受到离群值污染，这些离群值可能由重大误差、故障状态、停电和极端天气等因素造成。所提出的模型会降低训练数据集测量值中的垂直离群值和不良杠杆点的权重。用于限制离群值影响的权重通过投影统计量计算得出；投影统计量是时间序列数据点马氏距离的鲁棒版本。所提出的方法在 IEEE 33 节点配电系统和一个深度整合可再生能源的真实世界不平衡 240 节点配电系统上进行了验证。仿真结果表明，所提出的鲁棒模型能够处理训练数据集中高达25%的离群值。**

索引词—**时序随机潮流；稳健过程建模；稳健马氏距离；广义最大似然估计量；异常值检测与识别。**

I. 引言

目前，电力系统的实时运行与控制操作依赖于状态估计器提供的信息。状态估计器会按固定时间间隔处理一组测量值，其中包括选定线路和母线上的有功功率、无功功率潮流、功率注入以及电压幅值。这些测量值由 SCADA测量值、相量测量单元（PMU）和智能电子设备（IED）等计量设备采集 [1]。它们有助于开展时间序列潮流分析，以预测负荷持续曲线，进而确定配电系统中的过载状态。众所周知，这些测量值常会受到离群值的影响。例如，在故障状态下，开关设备中的涌流干扰会暂时导致测量值出现误差。配电系统中使用的通信方式常会受到强烈的电磁干扰，导致数据损坏 [2]。此外，PMU的异步采样时间 [3], [4] 以及电压互感器和电流互感器中的磁饱和与磁滞会导致电流和电压相量出现测量误差 [5]。在这些情况下，基于加权最小二乘法的状态估计会受到遮蔽效应和涂抹效应的影响，从而产生不准确的结果。由于可再生能源（RES）和分布式发电（DGs）引入了随机动态，随着其渗透率不断提高，这一问题会更加严重{。对于大型电力系统，要通过数千次仿真实现不确定性量化，进行经典蒙特卡洛（MC）模拟需要很强的计算能力。因此，开发能够处理实际测量值、分析电力系统随机动态的稳健且计算高效的模型与工具至关重要。

文献中已提出若干随机潮流方法，用于开展灵敏度分析和不确定性量化 [6]。其中，最常用的方法是蒙特卡洛模拟和元模型。如前所述，当不确定性量化需要数千次仿真才能获得有意义的统计结果时，蒙特卡洛方法的计算效率并不理想。

元模型，也称为模拟器、代理模型或响应面，仅对确定性潮流仿真器进行统计表示。基于高斯过程的元模型属于非参数降阶模型，其模型输出样本被假定服从高斯分布 [7]。文献中广泛开发的其他类型元模型则基于多项式混沌 [8]–[15]。Ni *et al*.[9] 开发了稀疏多项式混沌展开，以处理大量随机输入变量并考虑这些变量之间的相关性。Wu *et al*. [10]提出了采用直角坐标形式的广义多项式混沌（gPC），以保留潮流的非线性。Wang *et al*. [11] 将 gPC 扩展为数据驱动广义多项式混沌，以更好地处理输入变量间的依赖相关不确定性。Xu *et al*. [12]开发了层次多项式混沌方差分析（ANOVA），使 gPC 能够高效扩展到大规模系统，同时避免维数灾难。一些采用神经网络的方法包括深度神经网络模型 [16], [17]、图卷积网络模型 [18], 和图神经网络 [19]，以克服计算方面的挑战。

元模型开发完成后，会在考虑数千个新的输入变量的情况下运行。该元模型在统计上代表潮流仿真器，可用于开展灵敏度分析和不确定性量化。 {然而，它们都不够稳健，无法基于实时数据进行训练，因此

本研究部分由美国国家科学基金会（NSF）1917308号项目以及中国东南大学科研启动基金3216002206A1号项目资助。P. Algikar和L. Mili任职于美国弗吉尼亚理工大学电气工程系，地址：美国弗吉尼亚州福尔斯彻奇，邮编22043。（电子邮箱：{apooja19, syarahmadi, lmili}@vt.edu）。Y. Xu任职于中国江苏省南京市东南大学。（电子邮箱：yijunxu@seu.edu.cn）。

不适合用于时序随机潮流分析。

元模型所得结果的准确性高度依赖训练数据的质量。要训练出良好的元模型，数据点必须覆盖整个输入设计空间。因此，需从随机输入变量设计空间中假定的概率分布采样数据点。通常假定负荷服从高斯分布、风速服从威布尔分布、太阳辐照度服从贝塔分布，等等。然而在实践中，这些分布可能无法代表实际数据 [20], [21]，从而导致不确定性量化结果不准确。传统的元建模方法无法处理功率曲线分布的失真，因此会产生有偏差的结果。

因此，文献中提出了现代数据驱动模型。随着传感与计量设备日益普及，这些模型利用原始观测数据，从而捕捉底层过程的自然随机性。例如，Wang 等在 [22]中利用多项式混沌开发了一种数据驱动仿真器，用于估计电压相量的统计特性；而 Xu 等在 [23] 中提出了一种完全非参数方法，以避免为高斯过程元模型假设参数分布。然而，所有这些方法都依赖原始数据，却未考虑离群值。众所周知，风力发电（WG）时间序列数据经常受到通信错误、风力涡轮机停机和限电的污染 [20] ；而光伏时间序列数据则会受到较大信号噪声、传感器故障、通信设备故障、最大功率跟踪异常、阵列停运和功率限制等因素的影响 [21]。

因此，人们提出了各种数据预处理技术来应对这些异常趋势。例如，Long *et al.* [24] 开发了一种算法，通过对风电功率曲线图像进行数学形态学运算，检测并清洗风力涡轮机异常数据。在 [25], 中，提出了一种方法，根据各个对象之间的相似程度，筛除原始风力数据中的离群值。随着电力系统规模不断扩大，在计算能力有限且新建风电场或太阳能电站的数据有限的情况下，这些数据清洗算法与非稳健的数据驱动模型（如 [22]，[23] ）相结合，会使实时随机分析耗时过长。它们是否适用于实时统计分析尚有争议。与本文提出的方法不同，这些方法未能与估计过程有效集成，也无法稳健地应对运行期间可能出现的所有类型的离群值。

本文基于稳健过程模型（RPM），提出一种实时数据驱动的时序随机潮流分析方法。所提出的 RPM 使用 Schweppe 型广义最大似然估计量（SHGM），能够处理训练数据集中高达 25% 的离群值。请注意，离群值可能是垂直离群值，也可能是不良高杠杆点。稳健统计中针对线性回归提出的其他估计量为 M 估计量，它们无法抵御不良高杠杆点 [26], [27]。在电力系统中，高杠杆点是指相对较短线路上的潮流，或接入线路相对较多的母线上的功率注入 [28]。我们仅通过三种统计概念从理论上评估 RPM 的稳健性，即影响函数、有限样本崩溃点和渐近最大偏差曲线。我们在辐射状 IEEE-33母线配电系统以及美国中西部一座实际的 240 母线配电系统上，通过向训练数据中人工添加最高达 25% 的垂直离群值和不良高杠杆点，实验验证了 RPM 的稳健性。

本文组织结构如下：第二节介绍传统高斯过程模型，随后阐述用于评估RPM稳健性的统计稳健性概念的理论背景。第三节介绍所提方法的构建过程。第四节展示RPM在IEEE 33母线配电系统和实际240节点配电系统中的性能。第五节总结全文并展望未来工作。

II. 背景

A. 高斯过程框架下时序随机潮流的建模

假设潮流仿真器 $f(\cdot)$ 是所有 p 母线在训练区间 $t_i \in \mathbb{R}^{2p}$ 内的有功功率和无功功率注入测量值的函数。输出变量为 $y_{t_i} = f(\mathbf{x}^{t_i}) + \epsilon_{t_i}$，表示某一母线的电压幅值或相角测量值，并带有独立同分布的加性高斯测量噪声 $\epsilon_{t_i} \sim \mathcal{N}(0, \sigma_0^2)$。在高斯过程建模框架中，潮流仿真器输出的不确定性由具有特定均值函数 $m(\cdot)$ 和协方差（核）函数 $k(\cdot,\cdot)$ 的高斯过程来描述。作为高斯过程分布，仿真器输出变量 $f(\mathbf{x}^{t_1})$ $f(\mathbf{x}^{t_2}) \ldots, f(\mathbf{x}^{t_n})$ 服从多元正态分布。为纳入我们对仿真器的先验认知，我们将先验分布建模为均值函数为 $m_0(\cdot)$、协方差函数为 $k_0(\cdot,\cdot)$ 的高斯过程。形式化地，有

$$f(\cdot)|\, \boldsymbol{\beta}, \boldsymbol{l}, \tau^2 \sim \mathrm{GP}(m_0(\cdot), k_0(\cdot,\cdot)), \tag{1}$$

其中，均值函数 $m_0(\cdot)$的形式为

$$m_0(\mathbf{x}) = \boldsymbol{h}(\mathbf{x})^T \boldsymbol{\beta}. \tag{2}$$

此处，$\boldsymbol{h}(\mathbf{x}^{t_i})$: $\mathbb{R}^{2p} \rightarrow \mathbb{R}^q$ 表示基函数，可用于对假定的电力系统非线性程度进行建模，即 $\boldsymbol{h}(\mathbf{x}^{t_i}) = [1, \mathbf{x}^{t_i}, \mathbf{x}_{t_i}^2, \mathbf{x}_{t_i}^3, \ldots]_{t_i}^T$。例如，常数、线性和二次基函数分别表示为$\boldsymbol{h}(\mathbf{x}) = [1, x_{t_{i1}}, \ldots, x_{t_{i2p}}, x_{t_{i1}}^2, \ldots, x_{t_{i2p}}^2]^T \in \mathbb{R}^q \quad q = 4p+1$)

$= [1]$, $\boldsymbol{h}(\mathbf{x}) = [1, x_{t_{i1}}, \ldots, x_{t_{i2p}}]^T$，以及 $\boldsymbol{h}(\mathbf{x})$ ， $i = 1\ 2\ \ldots\ n$ $k_0\ \ t_i\ \ t_j$ ， ， ，。核函数 $(\mathbf{x},\ \mathbf{x})$ 表示相应输出点之间的协方差（ y_{t_i}， y_{t_j} ）。一种常用的协方差函数是径向基函数，其定义为

$$k_0(\mathbf{x}_{t_i}, \mathbf{x}_{t_j} | \boldsymbol{l}) = \tau^2 \exp\left(-\sum_{k=1}^{2p} \frac{(\mathrm{x}_{t_{ik}} - \mathrm{x}_{t_{jk}})^2}{2lk^2} \right), \tag{3}$$

其中，$\boldsymbol{l}=(l_1,\ \ldots,\ l_{2p})$ 表示特征长度尺度，用于描述过程的变化快慢；$i,\ j=1,\ \ldots,\ n$。表 I 列出了其他一些协方差函数。

我们将 n 输入和输出测量值分别汇集到矩阵 $\mathbf{X}=[\mathbf{x}^{t_1},\ldots,\mathbf{x}_{t_n}]^T$和向量 $\mathbf{y}=[y_{t_1},\ldots,y_{t_n}]^T$中。基函数矩阵表示为 $\mathbf{H(X)}=[\boldsymbol{h}(\mathbf{x}^{t_1})\ldots,\boldsymbol{h}(\mathbf{x}^{t_n})]^T$。根据（1），电力系统输出向量 $\mathbf{y}$ 服从多元正态随机向量分布，其协方差函数为对角项与噪声项 $\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma_n^2\mathbf{I}_n)$的加和。形式上，我们有

$$\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\boldsymbol{l},\tau^2,\sigma_n^2\sim\mathcal{N}\left(\mathbf{H}(\mathbf{X})\boldsymbol{\beta},\boldsymbol{\Sigma}(\mathbf{X})\right),\tag{4}$$

其中 $\boldsymbol{\Sigma}(\mathbf{X})=\boldsymbol{k}_0(\mathbf{X},\mathbf{X})+\sigma_n^2\mathbf{I}_n$。均值为零、方差为 σ_n^2的噪声项 $\boldsymbol{\epsilon}$ 也称为“块金效应”，用于反映模型不确定性并确保数值稳定性。训练数据集由 $(\mathbf{y},\mathbf{X})$ 构成。

对于随机潮流分析，假设我们在预测区间内的各个时刻$\boldsymbol{t^*}=[1,\ldots,n^*]$ 抽取 K 个输入测试变量样本$\mathbf{x}_{t^*}$。在时间序列分析中，通常会假设某一特定时间范围内具有平稳性和遍历性。后者意味着样本平均（通常称为集合平均）等于时间平均。遍历性假设使我们能够使用每个时刻的一次实时测量来对随机时间序列潮流建模。将时刻 t_i^*的 K 个采样测试预测变量分组，记为$\mathbf{x}_{t_i^*}^{(k)}$，其中k取值于 $[1,\ K]$，组成$\mathbf{X}_i^*=[\mathbf{x}_{t_i^*}^{(1)},\ \ldots,\mathbf{x}_{t_i^*}^{(K)}]^T$。采用分层形式，通过潮流仿真器 $f(\cdot)$在测试点$\mathbf{X}_i^*$ 获得的模型输出变量$\mathbf{y}^*$与训练输出变量共同服从如下联合多元高斯分布：

$$\begin{bmatrix}\mathbf{y}\\ \mathbf{y}^*|\mathbf{X}_i^*\end{bmatrix}\sim\mathcal{N}\left(\begin{bmatrix}\boldsymbol{m}_0(\mathbf{X})\\ \boldsymbol{m}_0(\mathbf{X}_i^*)\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma}(\mathbf{X}) & \mathbf{C}(\mathbf{X}_i^*)\\ \mathbf{C}^T(\mathbf{X}_i^*) & \mathbf{V}(\mathbf{X}_i^*)\end{bmatrix}\right),\tag{5}$$

其中，$\mathbf{C}(\mathbf{X}_i^*)=\boldsymbol{k}_0(\mathbf{X},\mathbf{X}_i^*),\mathbf{C}^T(\mathbf{X}_i^*)=\boldsymbol{k}_0(\mathbf{X}_i^*,\mathbf{X})$ 和 $\mathbf{V}(\mathbf{X}_i^*)=\boldsymbol{k}_0(\mathbf{X}_i^*,\mathbf{X}_i^*)$。协方差矩阵 $\boldsymbol{\Sigma}(\mathbf{X})$ 以下记作 $\boldsymbol{\Sigma}$。此外，我们假设仿真器在测试点 $f(\mathbf{X}_i^*)|\mathbf{X}_i^*$ 处的输出服从先验高斯概率分布，即

$$f(\mathbf{X}_i^*)|\mathbf{X}_i^*\sim\mathrm{GP}\left(\boldsymbol{m}_0(\mathbf{X}^*),\mathbf{V}(\mathbf{X}_i^*)\right).\tag{6}$$

进行条件化并使用多元分布的标准方法后，可得

$$f(\mathbf{X}_i^*)|\mathbf{X}_i^*,\mathbf{y},\mathbf{X},\boldsymbol{\beta},\boldsymbol{l},\tau^2,\sigma_n^2\sim\mathrm{GP}\left(\boldsymbol{\mu}^*(\mathbf{X}),\boldsymbol{\Sigma}^*(\mathbf{X})\right),\tag{7}$$

其中，估计均值函数 $\widehat{\boldsymbol{\mu}}^*(\mathbf{X}_i^*)$ 为

$$\widehat{\boldsymbol{\mu}}^*(\mathbf{X}_i^*)=\widehat{\boldsymbol{m}}_0(\mathbf{X}_i^*)+\widehat{\mathbf{C}}^T(\mathbf{X}_i^*)\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{r},\tag{8}$$

且估计协方差函数 $\boldsymbol{\Sigma}^*(\mathbf{X}^*)$ 表示为

$$\widehat{\boldsymbol{\Sigma}}^*(\mathbf{X}_i^*)=\widehat{\mathbf{V}}(\mathbf{X}_i^*)-\widehat{\mathbf{C}}^T(\mathbf{X}_i^*)\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\mathbf{C}}(\mathbf{X}_i^*),\tag{9}$$

对于预测区间内的所有实例，$i=1,\ \ldots,\ n^*$。(8) 给出的均值函数估计是一种计算高效的代理模型，可捕捉潮流的行为；而 (9) 给出的协方差矩阵估计则量化了相关的不确定性。

令 $(\boldsymbol{\beta},\ \tau^2)$ 服从弱先验 $p(\boldsymbol{\beta},\ \tau^2)\ \propto\ \frac{1}{\tau^2}$，结合 (4) 并应用贝叶斯定理，可得 $(\boldsymbol{\beta},\ \tau^2)$ 的后验分布，该分布为正态逆伽马分布，表示为

$$\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\boldsymbol{l},\tau^2,\sigma_n^2\sim\mathcal{N}(\hat{\boldsymbol{\beta}},\tau^2(\mathbf{H}^T\boldsymbol{\Sigma}^{-1}\mathbf{H})^{-1}),\tag{10}$$

其中，$\boldsymbol{\beta}$ 是加权最小二乘估计，表示为 $\boldsymbol{\beta}=(\mathbf{H}^T\boldsymbol{\Sigma}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\boldsymbol{\Sigma}^{-1}$y</style}，且

$$\tau^2|\mathbf{y},\mathbf{X},\boldsymbol{l},\sigma_n^2\sim\mathrm{InvGamma}\left(\frac{n-q}{2},\frac{(n-q-2)\hat{\tau}^2}{2}\right),\tag{11}$$

where $\hat{\tau}^2=\frac{\mathbf{y}^T(\boldsymbol{\Sigma}^{-1}-\boldsymbol{\Sigma}^{-1}\mathbf{H}(\mathbf{H}^T\boldsymbol{\Sigma}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\boldsymbol{\Sigma}^{-1})\mathbf{y}}{(n-q-2)}$.

B. 传统 *GPM* 设计中的涂抹效应与遮蔽效应

残差向量定义为观测向量$\mathbf{y}$与估计向量 $\widehat{\mathbf{y}}$之差。形式上，有

$$\boldsymbol{r}=\mathbf{y}-\widehat{\mathbf{y}},\tag{12}$$

其中

$$\hat{\mathbf{y}}=\mathbf{S}\mathbf{y},\tag{13}$$

其中$\mathbf{S}$ 为帽子矩阵，定义为$\mathbf{S}=\mathbf{H}(\mathbf{H}^T\boldsymbol{\Sigma}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\boldsymbol{\Sigma}^{-1}$. 将式（13）代入式（12），得到

$$\boldsymbol{r}=\mathbf{y}-\mathbf{S}\mathbf{y},\tag{14}$$

$$=(\mathbf{I}-\mathbf{S})\mathbf{y},\tag{15}$$

$$=\mathbf{W}\mathbf{y}.\tag{16}$$

将表达式$\mathbf{y}=\boldsymbol{m}(\mathbf{X})+\boldsymbol{e};\ \ \boldsymbol{e}\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})$ 代入式（16），得到

$$\boldsymbol{r}=\mathbf{W}\boldsymbol{e}.\tag{17}$$

此处，$\mathbf{W}$ 称为残差敏感度矩阵，因为它表示残差对误差的敏感度。如下所示，对于加权最小二乘估计量，离群值检测与识别的统计检验会受到离群值对残差的涂抹效应和掩蔽效应的影响。

- 涂抹效应
  假设对于 $i=2,\ 3,\ \ldots\ldots,\ n,\ e_1\neq0$ 且 $e_i=0$。由（14）式可得

$$r_1=W_{11}e_1\neq0,\tag{18}$$

$$r_2=W_{21}e_1\neq0,\tag{19}$$

$$r_n=W_{n1}e_1\neq0.\tag{20}$$

  这称为单个离群值对残差的涂抹效应，使得难以通过残差统计检验识别该离群值。

- 掩蔽效应
  假设对于 $i=3,\ 4,\ \ldots\ldots,\ n,\ e_1\neq0$ 且 $e_2\neq0$，而 $e_i=0$。相应的残差表示为

$$r_1=W_{11}e_1+W_{12}e_2,\tag{21}$$

$$r_2=W_{21}e_1+W_{22}e_2.\tag{22}$$

  因此，存在 e_1 和 e_2，使得残差 $r_1\approx0$ 且 $r_2\approx0$。这称为离群值对残差的掩蔽效应，会导致离群值残差识别检验失效。

表 I：常用核函数

Type	表达式
指数核 $k_E(\mathbf{x}^{t_i}, \mathbf{x}^{t_j})$	$\tau^2 \exp\left(-\sum_{k=1}^{2p} \frac{\lvert \mathrm{x}^{t_{ik}} - \mathrm{x}^{t_{jk}} \rvert}{l_k}\right)$
Matérn核 $3/2$ $k_M(\mathbf{x}^{t_i}, \mathbf{x}^{t_j})$	$\tau^2 \left(1+\sum_{k=1}^{2p} \frac{\sqrt{3}\lvert \mathrm{x}^{t_{ik}} - \mathrm{x}^{t_{jk}} \rvert}{l_k}\right) \exp\left(-\sum_{k=1}^{2p} \frac{\sqrt{3}\lvert \mathrm{x}^{t_{ik}} - \mathrm{x}^{t_{jk}} \rvert}{l_k}\right)$
有理二次 $k_{RQ}(\mathbf{x}^{t_i}, \mathbf{x}^{t_j})$	$\left(1+\exp\left(-\sum_{k=1}^{2p} \frac{(\mathrm{x}^{t_{ik}} - \mathrm{x}^{t_{jk}})^2}{2l_k^2 \alpha}\right)\right)^{-\alpha}$

C. 鲁棒性概念

本小节简要回顾估计量统计效率的定义，以及稳健统计中提出的鲁棒性概念，即渐近影响函数、崩溃点和渐近最大偏差曲线。

1) 渐近影响函数：对于一个 ϵ污染模型 $G(r) = (1-\epsilon)\Phi(r)+\epsilon F(r)$，其中 Φ 是累积高斯分布函数， F 是未知的残差分布函数。影响函数基于盖托导数，用于量化估计量 $T(G)$对任意无穷小污染$H = \Delta_{\boldsymbol{r}}$的局部敏感性。其表达式为

$$\mathrm{IF}(r_i, \boldsymbol{h}_i; \Phi) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)\Phi + \epsilon \Delta_r) - T(\Phi)}{\epsilon}. \quad (23)$$

线性回归模型的M估计量的总影响函数 IF(r; Φ)，等于残差的标量值影响 IR(r_i; Φ)与位置的向量值影响 IP($\boldsymbol{h}_i$; Φ)的乘积。形式上，有

$$\mathrm{IF}(r; \Phi) = \mathrm{IR}(r_i; \Phi)\mathrm{IP}(\boldsymbol{h}_i; \Phi). \quad (24)$$

它们分别定义为 IR(r_i; Φ) $= \frac{\psi(\frac{r_i}{s})}{\mathcal{E}[\psi'(\frac{r_i}{s})]}$ 和 IP($\boldsymbol{h}_i$; Φ) $= (\mathbf{H}^T\mathbf{H})^{-1}\boldsymbol{h}_i$。对于 M估计量，当$\psi(\cdot)$有界时，IR(·)有界；而 IP(·)始终无界，这表明它对不良高杠杆点不具鲁棒性。

2）有限样本崩溃点：使估计量的最大偏差有限的ϵ最大值（记为 ϵ^*）称为该估计量的有限样本崩溃点。形式上，我们有 $\epsilon^* = max\{\epsilon; b_{max}(\epsilon) < \infty\}$。[29]和 [30]证明，在一般位置假设下，任何回归等变估计量的有限样本崩溃点最大值为 $\lfloor \frac{n-q}{2} \rfloor / n$ [31]。

3）渐近最大偏差曲线：渐近最大偏差曲线表示随着污染水平增加，估计量偏差上界所形成的曲线$0 \le \epsilon < \epsilon^*$。对于任何费 Fisher 信息一致的估计量 $\boldsymbol{\theta}$，其泛函形式为 $\boldsymbol{T}$，在任意 ϵ污染模型下的渐近最大偏差曲线定义为 $b_{max}(\epsilon) = \sup|\boldsymbol{T}(G) - \boldsymbol{\theta}|$。在$H$位置情形下，Huber [30]证明，样本中位数在所有位置等变估计量中具有可能的最小渐近最大偏差曲线。在线性回归中，具有最小渐近偏差曲线的估计量尚不明确。

4）统计效率：在假设的概率分布 F下，任何位置估计量 $\boldsymbol{\theta}_n$所能达到的最小方差ˆ由 Cramér–Rao 下界给出，该下界定义为费 Fisher 信息 I_f的倒数。形式上，我们有

$$Var(\sqrt{n}\hat{\boldsymbol{\theta}}_n; F) \ge \frac{1}{I_f}; \ \forall n. \quad (25)$$

克拉美-罗下界与估计量方差之比称为该估计量的效率。当 n 趋于无穷大时，渐近有效估计量的方差将达到克拉美-罗下界。

III. 稳健数据驱动过程仿真器

本节讨论所提出的 RPM 模型的开发。我们将式 (4) 重写为关于均值函数超参数的回归问题，其形式如下：

$$\mathbf{y}(\mathbf{X}) = \mathbf{H}(\mathbf{X})\boldsymbol{\beta} + \boldsymbol{e}, \quad (26)$$

通过求解下式，可获得采用最大似然估计方法的均值函数和协方差函数的超参数：

$$\widehat{\boldsymbol{\theta}} = \underset{(\boldsymbol{\beta}, \boldsymbol{l}, \tau^2, \sigma_n^2) \in \mathbb{R}^q \mathbb{R}^{2p} \mathbb{R}^{+*} \mathbb{R}^{+*}}{\arg\max} \log \mathcal{L}\left(\mathbf{y} | \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{l}, \tau^2, \sigma_n^2\right), \quad (27)$$

其中， $\boldsymbol{\theta}$表示一个超参数向量（$\boldsymbol{\beta}$， $\boldsymbol{l}$， τ^2， σ_n^2）。

A. Schweppe 型广义最大似然估计量

我们提出使用 SHGM估计器以稳健的方式估计 $\boldsymbol{\beta}$。SHGM估计器通过最小化残差 r_i 的加权损失函数来进行估计，该函数定义为

$$J(\boldsymbol{\beta}) = \min_{\hat{\boldsymbol{\beta}}} \sum_{i=1}^{n} w_i^2 \rho\left(\frac{r_i}{w_i s}\right), \quad (28)$$

其中， $\rho(\cdot)$ 是标准化残差 $r_{Si} = \frac{r_i}{w_i s}$ 的非线性损失函数。当对误差协方差知之甚少或一无所知时，残差尺度 s 通过 $s^* = 1.4826\ b_m$ 中位数 $|\boldsymbol{r}|$ 进行稳健估计。

$$s = 1.48261 + \frac{5}{n-q}\ \mathrm{median}|\boldsymbol{r}|. \quad (29)$$

我们选择 Huber 损失函数，是因为它具有凸性，且在中心区域呈二次特性。其定义如下：

$$\rho(r_i) = \begin{cases} \frac{r_i^2}{2} & \text{for } r_i < c, \\ c|r_i| - \frac{c^2}{2} & \text{for } r_i \ge c. \end{cases} \quad (30)$$

阈值参数 c 通常设为 1。5，这样可以在高斯分布下的高统计效率与抵抗离群值的良好鲁棒性之间取得良好平衡。

B. 基于投影统计量的权重

权重根据投影统计量计算得出，投影统计量是马氏距离的一种稳健形式。形式上，它们定义为：将点云沿着从逐坐标中位数出发并经过每个数据点的方向进行投影后，所得标准化投影距离的最大值 [28])。令 $\boldsymbol{h}^T(\mathbf{x}_i)$ 表示为 $\boldsymbol{h}_i^T$。形式上，有

$$\mathrm{PS}_i = \max_{||\boldsymbol{v}||=1} \frac{\boldsymbol{h}_i^T\boldsymbol{v} - \underset{j}{\mathrm{med}}(\boldsymbol{h}_j^T\boldsymbol{v})}{1.4826 \underset{k}{\mathrm{med}} |\boldsymbol{h}_k^T\boldsymbol{v} - \underset{j}{\mathrm{med}}(\boldsymbol{h}_j^T\boldsymbol{v})|}, \tag{31}$$

其中 $\boldsymbol{v}_j = \frac{\boldsymbol{u}_j}{||\boldsymbol{u}_j||}$; $\boldsymbol{u}_j = \boldsymbol{h}_i - \mathbf{M}$; $j = 1, \ldots, n$。这里，$\mathbf{M}$ 表示逐坐标中位数，其定义为

$$\mathbf{M} = \{ \underset{j=1,\ldots,n}{\mathrm{med}} \boldsymbol{h}_{j1}, \ldots, \underset{j=1,\ldots,n}{\mathrm{med}} \boldsymbol{h}_{jq}\}.$$

权重计算如下：

$$w(\boldsymbol{h}_i) = \begin{cases} 1, & \mathrm{PS}_i^2 \le b; \\ \frac{b}{\mathrm{PS}_i^2}, & \text{otherwise.} \end{cases} \tag{32}$$

当数据点 $\boldsymbol{h}_i$ 对应的 PS_i^2 大于 b 时，该数据点被视为高杠杆点。这些权重会降低不良高杠杆点和垂直离群值的权重，同时保留好杠杆点。

C. 均值函数超参数的稳健估计

我们通过令目标函数 $J(\boldsymbol{\beta})$ 关于 $\boldsymbol{\beta}$ 的梯度为零来估计均值的超参数，得

$$\sum_{i=1}^{n} w_i \boldsymbol{h}_i \frac{\partial \rho(r_{Si})}{\partial r_{Si}} = 0. \tag{33}$$

令 psi函数定义为 $\psi(r_{Si}) = {}^{\partial \rho_{(r_{Si})}}/_{\partial r_{Si}}$。式（33）现在变为

$$\sum_{i=1}^{n} w_i \boldsymbol{h}_i \psi(r_{Si}). \tag{34}$$

将式（34）除以标准化残差 r_{S_i}，得到

$$\sum_{i=1}^{m} q\left(\frac{r_i}{w_i s}\right) \boldsymbol{h}_i r_i = \mathbf{0}, \tag{35}$$

其中，$q(r_{Si}) = {}^{\psi_{(r_{Si})}}/_{r_{Si}}$ 称为权重函数。对于 Huber ρ-函数，其定义为

$$\boldsymbol{q}(r_{S_i}) = \begin{cases} 1, & r_i \le c \\ \frac{b\ \mathrm{sign}(r_{S_i})}{r_{Si}}, & \text{otherwise} \end{cases}. \tag{36}$$

代入残差的表达式 r_i，并将 (36) 改写为矩阵形式，得到

$$\mathbf{H}^T\mathbf{Q}(\mathbf{y} - \mathbf{H}\boldsymbol{\beta}) = \mathbf{0} \tag{37}$$

$$\boldsymbol{\beta} = (\mathbf{H}^T\mathbf{Q}\boldsymbol{\Sigma}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{Q}^{(k)}\boldsymbol{\Sigma}^{-1}\mathbf{y}, \tag{38}$$

其中 $\mathbf{Q}$ = 对角矩阵($q(r_{S_i})$)。由于 $\boldsymbol{\beta}$ 是 $\mathbf{Q}$ 的函数，我们通过采用迭代重加权最小二乘（IRLS）算法，以迭代方式求解 $\boldsymbol{\beta}$。形式化地说，我们有

$$\boldsymbol{\beta}^{(k+1)} = (\mathbf{H}^T\mathbf{Q}^{(k)}\boldsymbol{\Sigma}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{Q}^{(k)}\boldsymbol{\Sigma}^{-1}\mathbf{y}. \tag{39}$$

D. 超参数估计的迭代过程

本小节讨论与协方差函数相关的 RPM 超参数（$\boldsymbol{l}$，τ^2，σ_n^2）的稳健估计。利用代码的蒙特卡洛模拟所获得的观测集（$\mathbf{y}$，$\mathbf{X}$），我们使用公式（39）给出的算法估计超参数 $\widehat{\ }\ \boldsymbol{\beta}$。其余超参数，即（$\boldsymbol{l}$，$\tau^2$，$\sigma_n^2$）的最大似然估计可表示为

$$(\widehat{\boldsymbol{l}}, \widehat{\tau}^2, \widehat{\sigma}_n^2) = \underset{\boldsymbol{l},\tau^2,\sigma_n^2}{\arg\max} \log L\left(\mathbf{Y}|\mathbf{X}, \widehat{\boldsymbol{\beta}}, \boldsymbol{l}, \tau^2, \sigma_n^2\right). \tag{40}$$

将得到的对数 L 函数定义为

$$\Gamma(\boldsymbol{l}, \tau^2, \sigma_n^2) = \log |\boldsymbol{k}(\mathbf{X}, \mathbf{X}|\boldsymbol{l}, \tau^2) + \sigma_n^2\mathbf{I}_n|. \tag{41}$$

因此，（$\boldsymbol{l}$，τ^2，σ_n^2）的最大似然估计可简化为

$$(\widehat{\boldsymbol{l}}, \widehat{\tau}^2, \widehat{\sigma}_n^2) = \underset{\boldsymbol{l},\tau^2,\sigma_n^2}{\arg\min} \Gamma(\boldsymbol{l}, \tau^2, \sigma_n^2). \tag{42}$$

利用 [32] 中所述的基于梯度的优化器估计超参数（$\widehat{\boldsymbol{l}}$，$\widehat{\tau}^2$，$\widehat{\sigma}_n^2$）。随后，我们可以将 $\widehat{\boldsymbol{\beta}}$ 更新为 $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}$（$\widehat{\boldsymbol{l}}$，$\widehat{\tau}^2$，$\widehat{\sigma}_n^2$）。所使用的算法用于

算法 1 用于估计 RDP 超参数的算法

1：在测量值不可用的情况下构建电力系统仿真器；2：在所需输入功率测量值下运行仿真器，获取每个负荷母线处的电压幅值和电压相角，构成训练数据集（$\mathbf{X}$，$\mathbf{y}$）；3：使用合适的基函数构造$\mathbf{H}$；4：根据式（31）计算$\mathbf{H}$各行向量的投影统计量；5：根据式（32）给出的 PS 计算权重 $\boldsymbol{w}$；6：使用加权最小二乘解初始化 $\boldsymbol{\beta}$，即 $\boldsymbol{\beta}_0 = (\mathbf{H}^T\boldsymbol{\Sigma}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\boldsymbol{\Sigma}^{-1}\mathbf{y}$；7：在将超参数（$\boldsymbol{l}$，$\tau^2$，$\sigma_n^2$）设为初始值的同时，执行式（39）给出的迭代重加权最小二乘算法，更新 $\boldsymbol{\beta}$ 直至收敛，得到 $\widehat{\boldsymbol{\beta}}$；8：在固定 $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$的同时，更新（$\boldsymbol{l}$，$\tau^2$，$\sigma_n^2$）；9：迭代执行步骤 7 和 8 直至收敛，例如 $||\boldsymbol{r}|| \le 0.001$，以获得最终的超参数估计值（$\widehat{\boldsymbol{\beta}}$，$\widehat{\boldsymbol{l}}$，$\widehat{\tau}^2\widehat{\sigma}_n^2$）。

超参数的估计过程总结于算法 1 中。一旦 RPM 的所有超参数（$\boldsymbol{\beta}$，$\boldsymbol{l}$，τ^2，σ_n^2）估计完成，我们便使用（8）作为一种稳健且计算高效的代理模型，并使用（9）量化其方差。

*E. RPM*的稳健性

SHGM估计器的影响函数为

$$\mathrm{IF}(r_{Si}, \boldsymbol{h}_i; \Phi) = \frac{\psi(r_{Si})}{E_\Phi[\psi'(r_{Si})]}(\mathbf{H}^T\mathbf{H})^{-1}\boldsymbol{h}_i w_i, \qquad (43)$$

对于 SHGM 估计器，可以注意到位置的影响 $\mathrm{IP}(\boldsymbol{h}_i, \Phi) = (\mathbf{H}^T\mathbf{H})^{-1}\boldsymbol{h}_i w_i$ 是有界的，这得益于利用投影统计量计算的权重（见第三节），其崩溃点达到 $\lfloor (n-q-1)/2 \rfloor / n$ [31] 所给出的最大值。请注意，对于较小的标准化残差，SHGM 估计器会退化为 ℓ_2 范数估计量；对于较大的标准化残差，则退化为 ℓ_1 范数估计量。因此，它在高斯分布下具有较高的统计效率，同时对离群值具有鲁棒性。

IV. 案例研究

本节将所提出的 RPM 模型与高斯过程模型（GPM）进行比较，考察它们应用于标准 IEEE 33节点系统（案例 A）以及位于美国中西部、可再生能源和分布式发电渗透率较高的实际 240节点配电系统（案例 B）时的性能。我们在训练数据中加入最多 25% 的垂直离群值（即 y 中的离群值）、不良杠杆点（即 X 中的离群值）和好杠杆点（即 (X, y) 中的离群值）。为展示 RPM 在非高斯分布噪声下的良好性能，我们假设噪声服从自由度为 10 的学生t分布。选择这种分布是因为其自由度较低时尾部较重，产生的采样值可能远离其中位数。对于每个案例，我们使用平均绝对误差和均方根指标比较 GPM 和 RPM 的性能。

*A. IEEE 33*节点系统

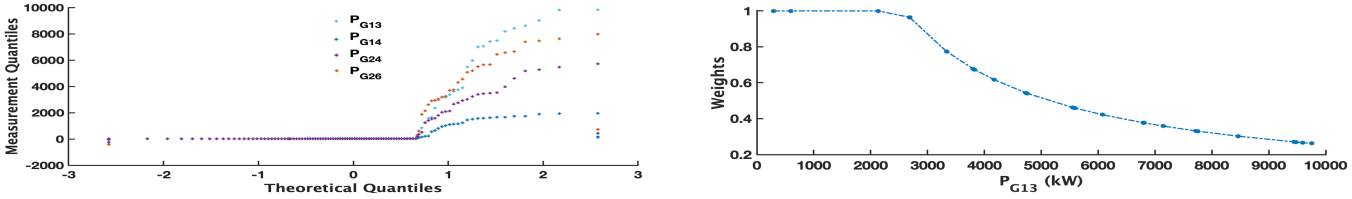
RPM应用于一个标准IEEE 33节点系统，该系统接入了四个可再生能源，即一套光伏（P_{G24}）接入24号母线，以及三台风力发电机（P_{G13}、P_{G14}、P_{G26}）分别接入13、14和26号母线，容量分别为1 kW、50 kW、10 kW和10 kW。所考虑的可再生能源出力和负荷时间序列数据为实际测量值，分辨率为 $1s$。我们在 $n = 150$ 个输入数据点 $\mathbf{X} = [\mathbf{x}_{t_1}、\ldots\ldots、\mathbf{x}_{t_{150}}]^T$ 处运行潮流仿真器，以获得相应的电压幅值和相角值 $\mathbf{y} = [y_{t_1}, \ldots, y_{t_{150}}]^T$，这些数据构成训练数据。基于（y，X）进行训练后，RPM 和传统高斯过程模型用于预测接下来的 $n^* = 60$ 个数据点，这些数据点在时刻 $\boldsymbol{t}^* = [t_{151}, \ldots, t_{210}]$构成验证数据集。为进行随机分析，我们采用拉丁超立方抽样，在验证数据集的每个时刻生成7000组输入变量样本；其中风力发电机遵循威布尔分布（$P_{t_i^*、G13}$、$P_{t_i^*、G14}$、$P_{t_i^*、G24} \sim$ Weibull(2.06, 7.1)），光伏遵循贝塔分布（$P_{t_i^*、G26} \sim$ Beta(2.06, 2.5)）；$i = 151、\ldots\ldots、210$。在这些样本上执行蒙特卡洛（MC）模拟所得的结果，作为与 RPM 和 GPM 结果进行比较的参考值。通过在训练数据集中加入25%的离群值（如图1（a）所示），展示 RPM 的鲁棒性。

具体而言，我们通过在输入数据点 $[\mathbf{x}_{t_1}、\ldots\ldots、\mathbf{x}_{t_{37}}]$中加入不良高杠杆点来构造最坏情况场景，即在测量值（$\mathbf{P}_{G13}$、$\mathbf{P}_{G14}$、$\mathbf{P}_{G24}$、$\mathbf{P}_{G26}$）以及负荷母线的负荷消耗 $\{\mathbf{P}_{L1}、\mathbf{P}_{L2}、\ldots\ldots、\mathbf{P}_{L33}\}$中加入此类点。同样，我们也在输出数据点 $[y_{t_1}, \ldots, y_{t_{37}}]$中加入垂直离群值，即在电压相量测量值中加入离群值。从图1（b）所示的权重可以看出，SHGM估计器降低了不良高杠杆点和垂直离群值的权重。训练数据中的离群值比例最高达25%时，19号母线电压幅值和相角的预测结果构成本研究的基准。图2（a）展示了19号母线电压相角预测结果的均值和标准差（以误差条表示），其中误差条表示标准差。图2（c）展示了 RPM 在训练时段 $[t_1 - t_{150}]$内得到的数据拟合结果。图2（b）将 RPM 根据下一时刻 t_{151}^* 的7000次实现计算得到的19号母线电压相角概率密度，与 MC 模拟输出进行了比较。图3（a）展示了 RPM 和 GPM 得到的19号母线电压幅值预测值。可以看到，GPM 的预测值与真实值相差很大。这是因为传统 GPM 的均值函数超参数估计以基本加权最小二乘估计为中心。因此，当存在离群值时，它无法准确表征仿真器，而 RPM 则能够做到。此外，图3（b）使用均方根误差（RMSE）值展示了训练数据集中离群值比例逐渐增加至25%时的预测精度。我们注意到，RPM 始终表现出较低的 RMSE 值。表II列出了训练数据集含有和不含离群值时，采用线性和二次基函数预测19号母线电压相量所得的 RMSE 和平均绝对误差（MAE）值。我们发现，在加入离群值的情况下，二次基函数的预测结果更准确。因此，根据简约原则，我们选择二次基函数来获得电网全部33条母线的电压相量结果，结果绘于图4。

B. 真实世界的*240*节点系统

我们将240节点辐射型配电系统 [33]与可再生能源相结合，即在电网中分布35台光伏发电机和35台风力发电机。它们在电网中的位置如图6所示。请注意，可再生能源接入电网中指定母线的每一相。

训练数据集（$\mathbf{X}$，$\mathbf{y}$）通过运行三相潮流仿真器获得，所用数据包括按小时采样的负荷以及连续 7 天的有功和无功功率注入测量值，即总计 $n = 168$ 个数据点构成用于设计 GPM 和 RPM 的训练数据集。GPM 和 RPM 均使用针对 $n^* = 24$ 个测试数据点（即验证数据点）获得的预测结果进行分析。对于概率分析，使用拉丁超立方抽样，从服从高斯分布的负荷输入变量中抽取 7000 个样本

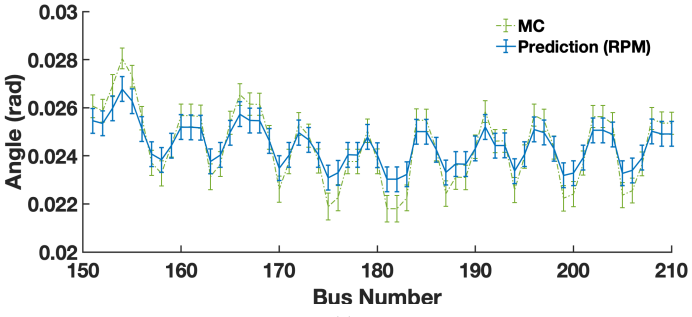


(a) (b)

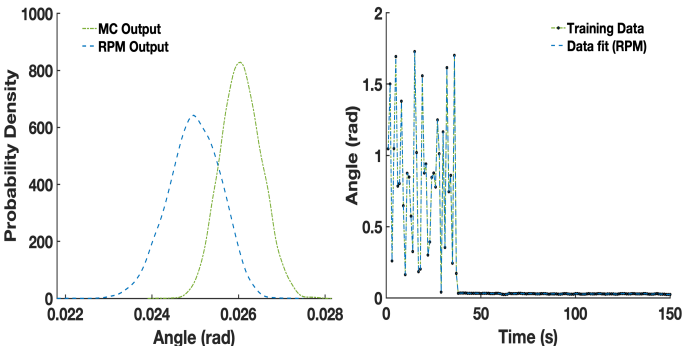
图1：离群值对训练数据集的污染；(a) 含有25%离群值的测量值的QQ图；(b) 使用PS与离群值幅值绘制的权重图。

表II：IEEE 33节点系统中19号母线的均方根误差和平均绝对误差

	二次基								线性基							
	含25%离群值				无离群值				含 25% 离群值				无离群值			
	RPM		GPM		RPM		GPM		RPM		GPM		RPM		GPM	
测量指标	V	A	V	A	V	A	V	A	V	A	V	A	V	A	V	A
RMSE	0.0034	$9.7810e^{-4}$	1.0274	0.1527	0.0931	0.0058	0.0657	0.1468	0.0264	0.01264	0.7995	0.0206	0.01005	0.01747	0.00838	0.01645
MAE	0.003	$8.2815e^{-4}$	7.3087	0.1287	0.0672	$2.5516e^{-4}$	0.0334	0.1669	0.0047	$9.7962e^{-4}$	4.2355	0.0029	0.00058	0.00176	0.00042	0.0018



(a)



(b) (c)

图 2：19号母线电压相角的 RPM 结果：(a) 测试点处的预测；(b) 测试点处的概率密度；(c) 训练数据点上的拟合值。

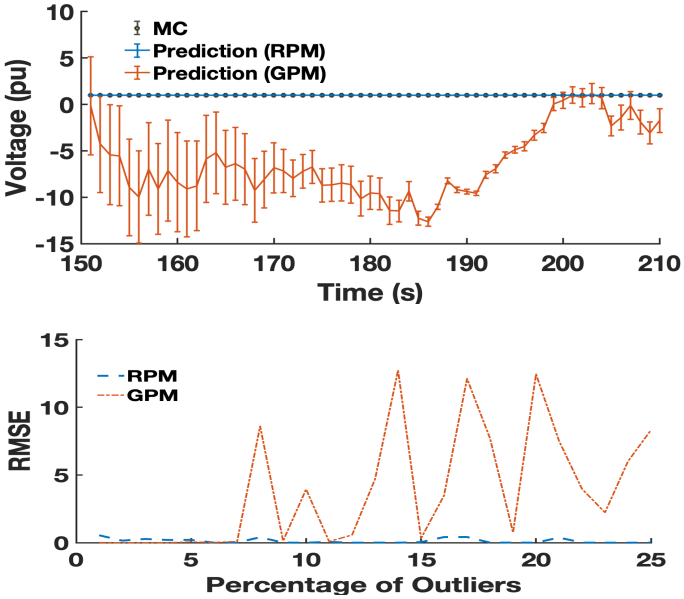
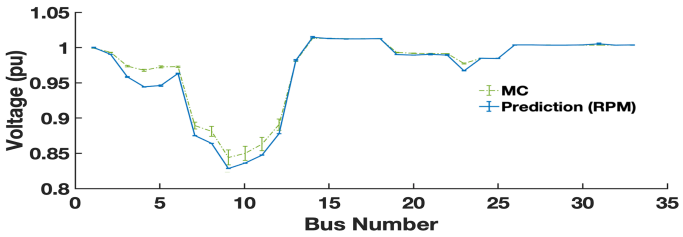


图 3：RPM 与 GPM 性能的比较：(a) 19号母线的电压幅值；(b) 均方根误差值。

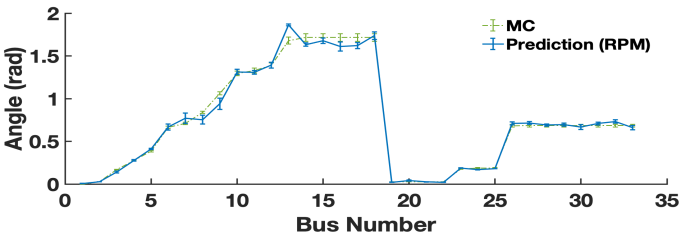
$P_{t_i^*,\,L} \sim \mathcal{N}(P_{t_i^*,\,L},\ 0.05P_{t_i^*,\,L})$，$i = 169,\ \ldots\ldots,\ 192$，风力发电机的输出服从威布尔分布，光伏发电机的输出服从贝塔分布，其形状参数和尺度参数与第 4.A 节中所述相同。将母线 2003。2 的电压相量预测作为日前预测，作为本研究的基准。我们在图 9(b) 和 (d) 中分别展示了使用线性和二次基函数时，GPM 和 RPM 得到的母线 2003。2 电压幅值预测结果的均值，其中误差条表示标准差，如其所示

分别绘制概率分布会很不方便。同样，Bus 2003。2 电压相角的预测结果如图 10(b) 和 (d) 所示。

现在，我们在功率注入测量值、电压幅值和相角中加入最多 25% 的离群值，即在训练数据集中前 42 个输入和输出数据点中加入离群值。为了充分展示 RPM 的鲁棒性，我们选择自由度为 10 的学生 t 分布作为离群值分布，因为如图 5 所示，该分布具有较厚的尾部。同样，在有功和无功负荷功率的测量值中也加入相同比例的离群值。图 7(b) 和 (d) 展示了采用线性和二次基函数时，测试输入在实例t_{169}^* 处得到的母线 2003.2 电压幅值的概率密度函数。类似地，母线 2003.2 电压角的概率密度结果

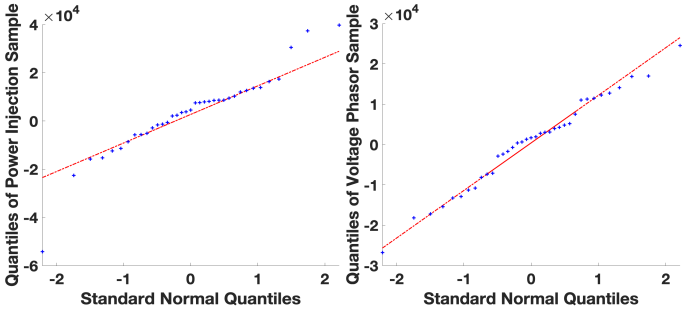


(a)



(b)

图4：训练数据集加入25%的离群值后，IEEE 33节点系统的RPM预测值：(a) 电压幅值；(b) 电压相角。



(a) (b)

图5：在(a) 功率注入测量值；(b) 电压幅值测量值中加入25%离群值后的QQ图。

绘于图 7(a) 和 (b)。图 9(a) 和 (c) 分别比较了采用线性和二次基函数时，由 GPM 和 RPM 得到的母线 2003.2 电压幅值预测结果。电压角结果见图 10(a) 和 (c)。训练数据集中加入离群值和未加入离群值两种情况下，采用线性基函数和二次基函数所得预测结果的均方根误差和平均绝对误差列于表 III。我们发现，对于电压幅值预测，在这两种情况下，GPM 和 RPM 采用线性基函数所得的均方根误差和平均绝对误差均小于采用二次基函数所得的误差。对于电压角，在训练数据集中含有离群值的情况下，两种模型采用二次基函数时表现更好；在不含离群值的情况下，线性基函数表现更好。因此，对于训练数据集中含有离群值的情况，我们选择线性基函数绘制电压幅值，并选择二次基函数绘制电压角，以进一步预测系统中所有母线的相关结果。均值（标示为

两种模型对240–母线系统各母线电压幅值（见图12）和相角（见图13）的预测结果（点）及标准差（以误差条表示）与蒙特卡洛模拟结果进行了比较。我们观察到，相应GPM得到的结果在均值和标准差方面均显著偏离真实值。由于SHGM估计器在准确性与鲁棒性之间进行了权衡，RPM的性能相对准确。传统高斯过程模型会强烈偏向离群值，因此其预测结果进一步偏离蒙特卡洛模拟结果。由于线路中的潮流较大，电压相量的方差较高，因此相 c 的结果中偏差尤为显著。对于幅值较大的离群值，由此产生的偏差是施加于电力系统测量值上的最坏情况场景。只要加入的离群值未超过崩溃点，所提出的RPM就能将该偏差保持在有限范围内；而传统高斯过程模型的结果偏差则是无界的。

训练数据中加入的离群值比例不断增加时，节点2003.2处电压幅值和相角预测值的均方根误差分别如图11(a)和(b)所示。我们观察到，GPM得到的均方根误差值高于RPM，而RPM始终能够得到较低的均方根误差结果。

备注1：请注意，由于无法获得电压相量（输出变量 $\mathbf{y}$）的实际测量值，因此使用实际测量的有功和无功功率注入（输入变量$\mathbf{X}$）运行潮流仿真器来获得这些电压相量。

V. 结论与未来工作

本文提出了一种稳健过程模型，利用功率注入和电压相量的时间序列测量值执行随机潮流计算。所提出的模型能够捕捉由可再生能源和分布式发电在电网中引入的自然随机动态。为此，我们使用包含可再生能源和分布式发电的有功和无功功率注入测量值以及节点电压相量的记录时间序列数据集对RPM进行训练。我们在标准IEEE 33节点系统和实际240节点系统上验证了RPM。结果表明，使用预测电压相量的均方根误差和最大绝对误差，所提出的方法能够处理训练数据集中25%的离群值，即不良杠杆点和垂直离群值。

未来的工作将着重拓展 RPM 在最优潮流计算中的应用。我们还将研究其他具有较高崩溃点的稳健估计量的性能，以处理训练数据中占比超过 25% 的离群值。

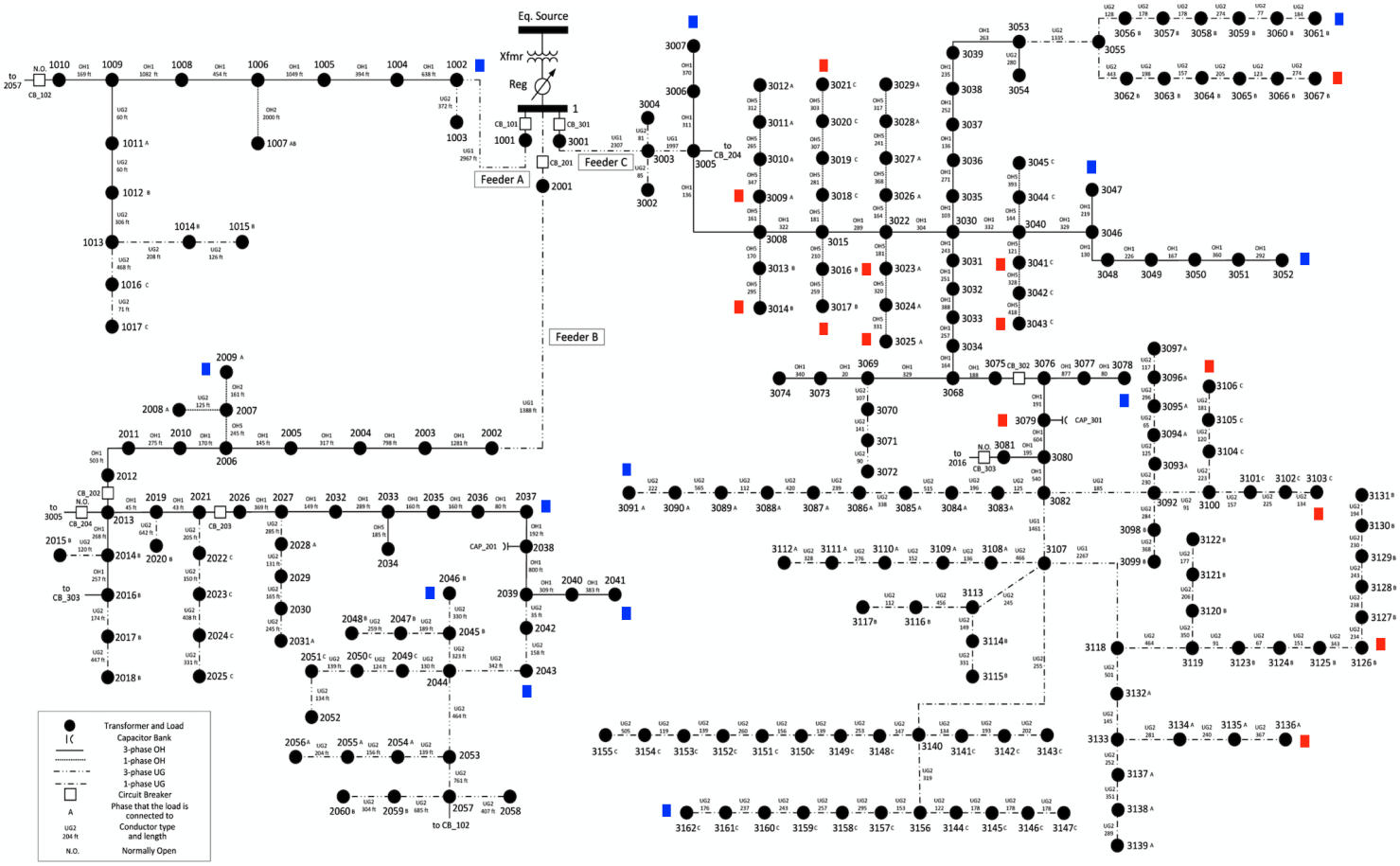


图6：集成可再生能源的240母线系统在线示意图。蓝色和红色方块分别表示光伏发电机和风力发电机

表III：240节点系统中母线2003.2的均方根误差和平均绝对误差

	二次基								线性基							
	含25%离群值				无离群值				含 25% 离群值				无离群值			
	RPM		GPM		RPM		GPM		RPM		GPM		RPM		GPM	
测量指标	*V*	*A*	*V*	*A*	*V*	*A*	*V*	*A*	*V*	*A*	*V*	*A*	*V*	*A*	*V*	*A*
RMSE	0.4222	3.8734	0.4242	5.5007	0.1888	4.9835	2.2246	11.1942	0.1921	8.8619	4.0814	53.4306	0.1183	4.4287	0.0637	6.1481
MAE	0.3845	3.1791	0.3164	4.5781	0.1544	3.8252	1.9426	8.5223	0.1859	3.3976	3.0779	34.3391	0.1026	1.5122	0.0513	4.9164

参考文献

[1] T. D. Liacco，“状态估计在电力系统运行中的作用”，*IFAC*会议论文集，第 15 卷，第 4 期，第 1531–1533 页，1982 年。[2] P. Zhang、F. Li、N. B. I. T. o. S. Grid 和 u. 2010，“面向未来智能控制中心的下一代监测、分析与控制”，*ieeexplore.ieee.org*。[3] A. Allen、S. Santoso 和 E. Muljadi，“ 用于筛查电力系统事件的相量测量单元数据的算法，以及相量测量单元相对相角差和频率信号中观测到的事件类别及其共同特征”，2013 年。[4] X. Liu、X. Zhang、L. Chen、F. X. J. o. M. P. ... 和 u. 2020，“通过马氏核回归和集成学习，考虑网络拓扑变化的数据驱动暂态稳定评估模型”，*ieeexplore.ieee.org*。[5] B. McCamish、R. Meier、J. Landford、R. B. E. P. S. ... 和 u. 2016，“用于高效管理电力系统测量值的后端框架”，*Elsevier*。[6] C. Lin、Z. Bie、B. Zhou、T. Wang 和 T. Wang，“相关风力发电条件下随机潮流的不同方法比较”，*IFAC-PapersOnLine*，第 51 卷，第 67–72 页，2018 年 1 月。[7] Y. Xu、Z. Hu、L. Mili、M. Korkali 和 X. Chen，“概率潮流基于高斯过程仿真器的潮流”，*IEEE*《电力系统汇刊》，第 35 卷，第 3278–3281 页，2020 年 7 月。[8] Z. Ren、W. Li、R. Billinton 和 W. Yan，“基于随机响应面法的概率潮流分析”，*IEEE*《电力系统汇刊》，第 31 卷，第 3 期，第 2307–2315 页，2016 年。[9] F. Ni、P. H. Nguyen 和 J. F. Cobben，“用于概率潮流的基自适应稀疏多项式混沌展开”，*IEEE*《电力系统汇刊》，第 32 卷，第 1 期，第 694–704 页，2016 年。[10] H. Wu、Y. Zhou、S. Dong 和 Y. Song，“基于广义多项式混沌的概率负荷潮流”，*IEEE*《电力系统汇刊》，第 32 卷，第 1 期，第 820–821 页，2016 年。[11] G. Wang、H. Xin、D. Wu、P. Ju 和 X. Jiang，“考虑相关不确定性的基于数据驱动任意多项式混沌的概率负荷潮流”，*IEEE*《电力系统汇刊》，第 34 卷，第 4 期，第 3274–3276 页，2019 年。[12] Y. Xu、L. Mili 和 J. Zhao，“基于分层自适应多项式混沌-方差分析方法的概率潮流计算与方差分析”，*IEEE*《电力系统汇刊》，第 34 卷，第 5 期，第 3316–3325 页，2019 年。[13] J. Laowanitwattana 和 S. Uatrongjit，“基于任意多项式混沌展开的含不确定可再生能源网络的概率潮流分析”，《电气与电子工程学会会刊》，第 13 卷，第 12 期，第 1754–1759 页，2018 年。

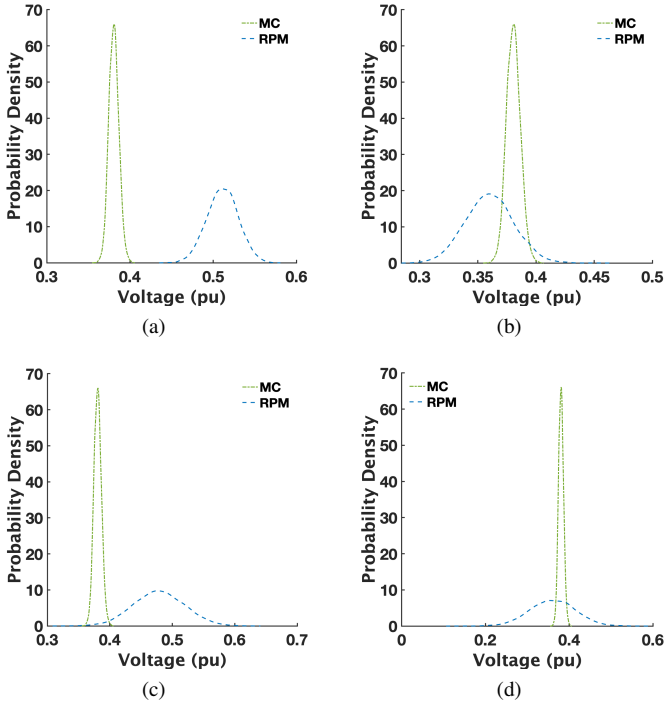


图7：在 240–母线电网中，母线2003.2的电压幅值概率密度结果对比：当（a）训练数据集加入25%的离群值；（b）线性基的训练数据集未加入离群值；（c）训练数据集加入25%的离群值；（d）二次基的训练数据集未加入离群值时，比较GPM与RPM的结果。

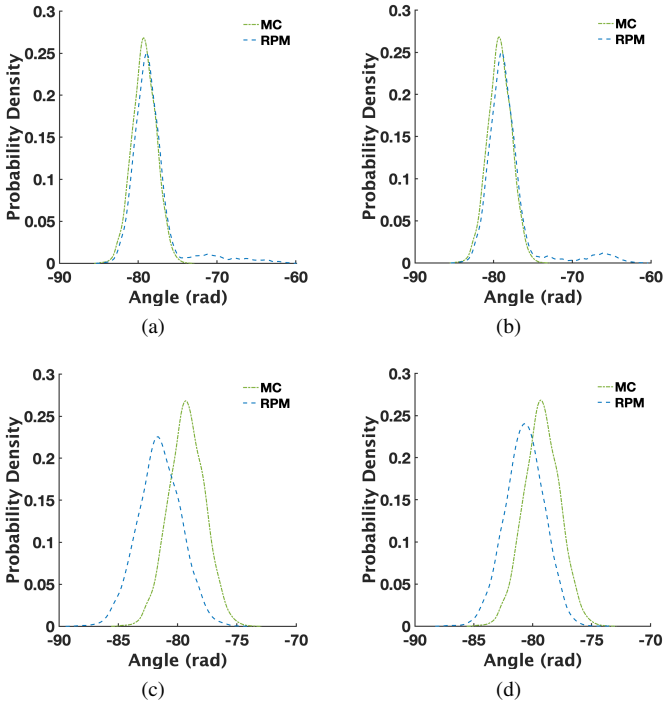


图8：在240–母线电网中，母线2003.2的电压相角概率密度结果对比：当（a）训练数据集加入25%的离群值；（b）线性基的训练数据集未加入离群值；（c）训练数据集加入25%的离群值；（d）二次基的训练数据集未加入离群值时，比较GPM与RPM的结果。

[14] J. Laowanitwattana 和 S. Uatrongjit，“基于偏最小二乘和任意多项式混沌展开的概率潮流分析”，*IEEE*《电力系统汇刊》，第 37 卷，1470，2021 年 2 月，第 1461 页。[15] K. Ye、J. Zhao、Y. Zhang、X. Liu 和 H. Zhang，“一种通用且计算高效的 Copula-多项式混沌框架，用于考虑光伏注入非线性相关性的概率潮流”，《国际电力与能源系统杂志》第 136 卷，文章编号 107727，2022 年。[16] Y. Yang、Z. Yang、J. Yu、B. Zhang、Y. Zhang 和 H. Yu，“概率潮流快速计算：一种基于模型的深度学习方法”，*IEEE*《智能电网汇刊》，第 11 卷，第 3 期，第 2235–2244 页，2019 年。[17] M. Xiang、J. Yu、Z. Yang、Y. Yang、H. Yu 和 H. He，“基于深度神经网络的拓扑变化下概率潮流分析”，《国际电力与能源系统杂志》，第 117 卷，文章编号 105650，2020 年。[18] D. Wang、K. Zheng、Q. Chen、G. Luo 和 X. Zhang，“基于图卷积网络的概率潮流求解”，载于*2020 IEEE PES* 欧洲创新智能电网技术会议（*ISGT-Europe*），第 650–654 页，IEEE，2020 年。[19] H. Wu、M. Wang、Z. Xu 和 Y. Jia，“基于图感知深度学习网络的配电系统概率潮流”，载于*2021 IEEE/IAS* 亚洲工业与商业电力系统会议（*I&CPS Asia*），第 105–109 页，IEEE，2021 年。[20] X. Ye、Z. Lu、Y. Qiao、Y. Min 和 M. O’Malley，“风电场时序功率数据中离群值的识别与校正”，*IEEE*《电力系统汇刊》，第 31 卷，第 4197–4205 页，2016 年 11 月。[21] G. Li、Z. Duan、L. Liang、H. Zhu、A. Hu、Q. Cui、B. Chen 和 W. Hu，“考虑光伏阵列输出分布特征的离群数据挖掘方法及其应用”，《能源报告》，第 6 卷，第 2345–2357 页，2020 年 11 月。[22] X. Wang、X. Wang、H. Sheng 和 X. Lin，“一种用于评估概率总传输的数据驱动稀疏多项式混沌展开方法

适用于含可再生能源电力系统的能力”，*IEEE*《电力系统汇刊》，第36卷，第3期，第2573–2583页，2020年。[23] Y. Xu、L. Mili、M. Korkali、K. Karra、Z. Zheng和X. Chen，“考虑风电渗透率的概率性负荷裕度评估数据驱动非参数方法”，*IEEE*《电力系统汇刊》，第35卷，第6期，第4756–4768页，2020年。[24] H. Long、L. Sang、Z. Wu和W. Gu，“基于风电功率曲线的图像式异常数据检测与清洗算法”，*IEEE Trans- actions on Sustainable Energy*，第11卷，第2期，第938–946页，2019年。[25] L. Zheng、W. Hu和Y. Min，“原始风电数据预处理：一种数据挖掘方法”，*IEEE Transactions on Sustainable Energy*，第6卷，第1期，第11–19页，2014年。[26] P. J. Rousseeuw和A. M. Leroy，稳健回归与离群值检测。John wiley & sons，2005年。[27] F. R. Hampel、E. M. Ronchetti、P. Rousseeuw和W. A. Stahel，稳健统计：基于影响函数的方法。Wiley-Interscience；纽约，1986年。[28] L. Mill、M. G. Cheniae和N. S. Vichare，“基于投影统计量的稳健状态估计”，*IEEE*《电力系统汇刊》，第11卷，第2期，第1118–1127页，1996年。[29] D. L. Donoho和M. Gasko，“基于半空间深度和投影离群度的位置估计的崩溃性质”，*The Annals of Statistics*，第1803–1827页，1992年。[30] P. J. Huber，“位置参数的稳健估计”，*Annals of Mathematical Statistics*，第35卷，第1期，第73–101页，1964年。[31] R. A. Maronna和V. J. Yohai，“Stahel-Donoho稳健多元估计量的行为”，*Journal of the American Statistical Asso- ciation*，第90卷，第429期，第330–341页，1995年。[32] C. Rasmussen和C. Williams，用于机器学习的高斯过程。自适应计算与机器学习，马萨诸塞州剑桥，美国：麻省理工学院出版社，2006年1月。[33] F. Bu、Y. Yuan、Z. Wang、K. Dehghanpour和A. Kimber，“基于实际公用事业数据的时间序列分布测试系统”，载于*2019*年北美电力研讨会（*NAPS*），第1–6页，IEEE，2019年。

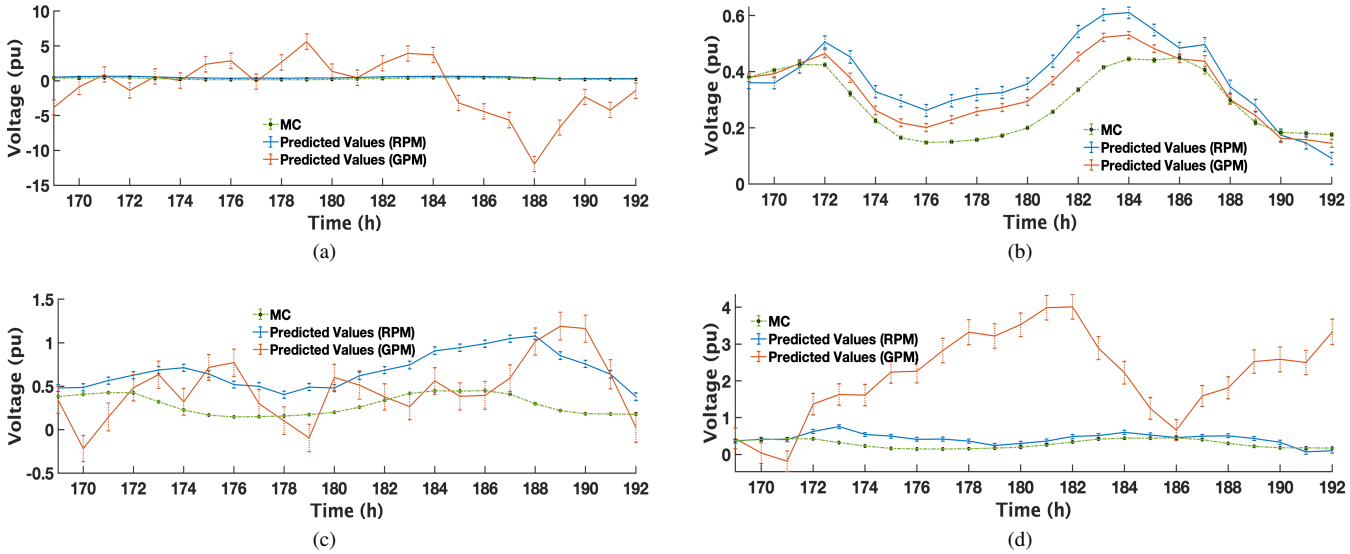


图9：当 240–母线电网中母线2003.2的电压幅值在以下情况下的 GPM 和 RPM 预测结果对比：(a) 训练数据集加入25%的离群值；(b) 线性基的训练数据集未加入离群值；(c) 训练数据集加入25%的离群值；(d) 二次基的训练数据集未加入离群值。

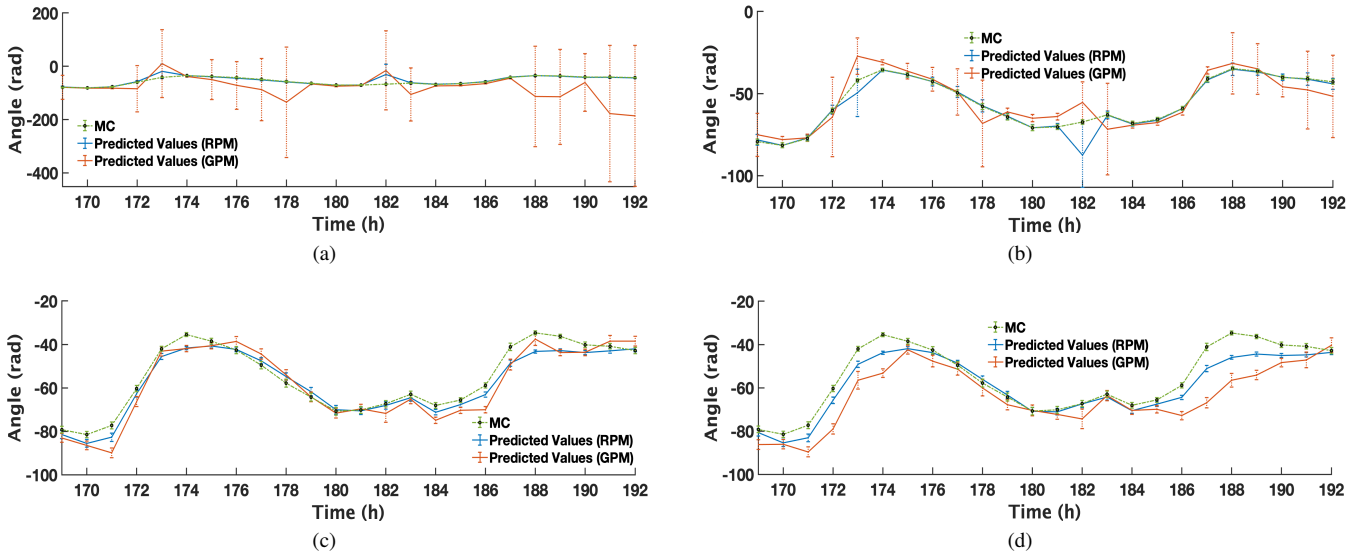


图10：当 240–母线电网中母线2003.2的电压相角在以下情况下的 GPM 和 RPM 预测结果对比：(a) 训练数据集加入25%的离群值；(b) 线性基的训练数据集未加入离群值；(c) 训练数据集加入25%的离群值；(d) 二次基的训练数据集未加入离群值。

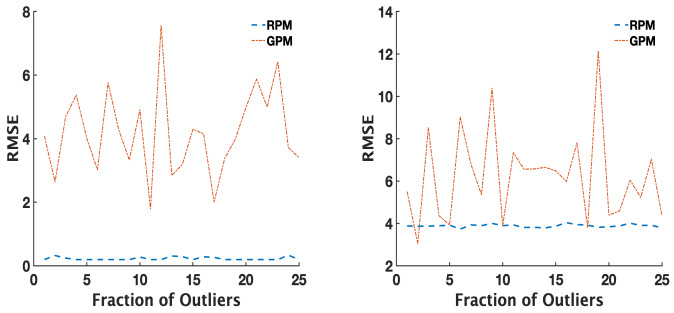


图11：训练数据中添加的离群值比例与母线 2003.2 处预测结果均方根误差的关系。(a) 电压幅值；(b) 电压角。

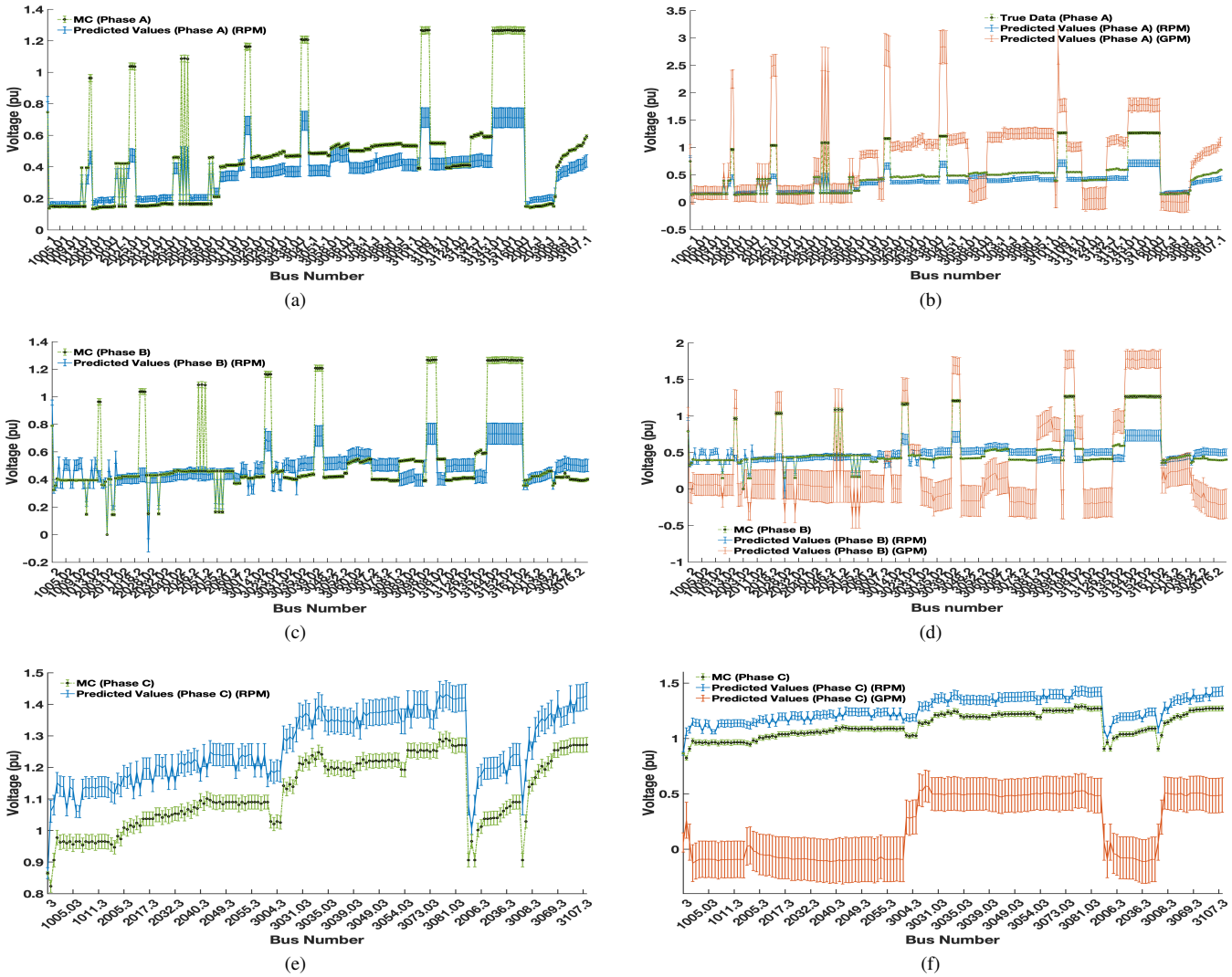


图12：训练数据中添加 25% 离群值时，RPM 得到的 240– 母线系统电压幅值预测结果与 GPM 得到的结果对比。(a) RPM 得到的 a 相结果；(b) GPM 与 RPM 的 a 相结果对比；(c) RPM 得到的 b 相结果；(d) GPM 与 RPM 的 b 相结果对比；(e) RPM 得到的 c 相结果；(f) GPM 与 RPM 的 c 相结果对比。

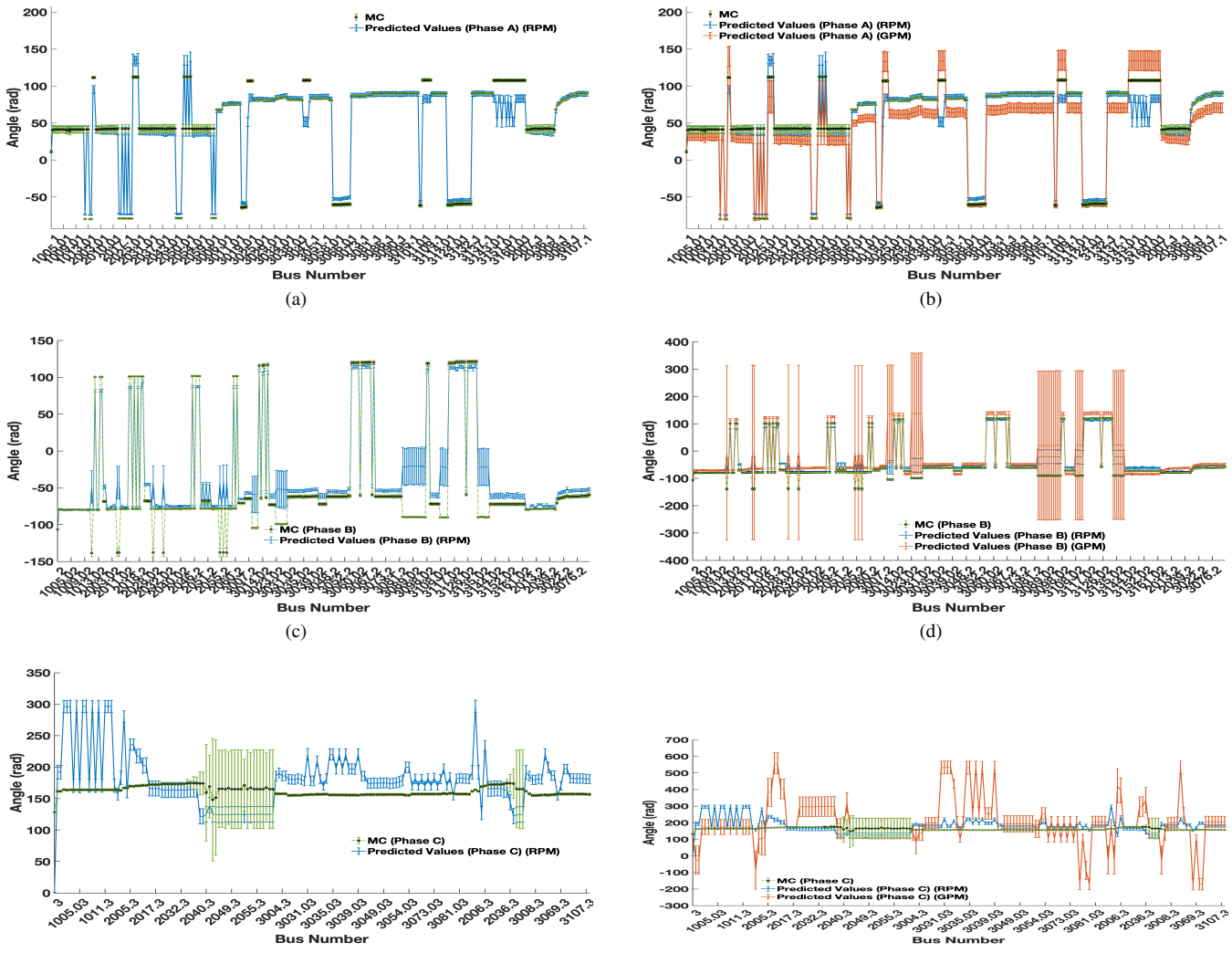


图13：在训练数据中加入 25% 的离群值后， 240–母线系统中 RPM 得到的电压相角预测结果与传统高斯过程模型得到的结果的比较。(a) RPM 得到的 a 相结果；(b) GPM 与 RPM 在 a 相上的比较；(c) RPM 得到的 b 相结果；(d) GPM 与 RPM 在 b 相上的比较；(e) RPM 得到的 c 相结果；(f) GPM 与 RPM 在 c 相上的比较。