Supplement to "Diagnostic Checking for Wasserstein Autoregression"

Chenxiao Dai[1], Feiyu Jiang[2], and Dong Li[3]

[1]Yau Mathematical Sciences Center, Tsinghua University
[2]Department of Statistics and Data Science, School of Management, Fudan University
[3]Department of Statistics and Data Science, Tsinghua University

The supplementary material contains the proofs of all the theorems presented in the main text, together with additional numerical and implementation details. Section S.1 proves the asymptotic result for the ATM(1) parameter estimator. Section S.2 proves the limiting distributions of the McLeod-type and sample-splitting portmanteau tests. Section S.3 gives additional discussion of the ATM(2) extension and the associated matrix cancellation issue. Section S.4 reports additional simulation experiments, and Section S.5 provides additional empirical prediction checks. Section S.6 examines the numerical verification of condition (4.13) across different maximum lag orders.

S.1 Proof of Theorem 3.1

Following Zhu and Müller (2023), we define the following quantities,

$$\widehat{\rho}_+^0 = \frac{1}{n-1}\sum_{i=2}^{n}\int_\Omega (T_{i-1}(x)-x)^2\mathrm{d}x, \quad \rho_+^0 = \int_\Omega \mathbb{E}[(T_1(x)-x)^2]\mathrm{d}x,$$
$$\widehat{\rho}_-^0 = \frac{1}{n-1}\sum_{i=2}^{n}\int_\Omega (x-T_{i-1}^{-1}(x))^2\mathrm{d}x, \quad \rho_-^0 = \int_\Omega \mathbb{E}[(x-T_1^{-1}(x))^2]\mathrm{d}x,$$

and

$$\widehat{\rho}_+^1 = \frac{1}{n-1}\sum_{i=2}^{n}\int_\Omega (T_i(x)-x)(T_{i-1}(x)-x)\mathrm{d}x, \quad \rho_+^1 = \int_\Omega \mathbb{E}[(T_2(x)-x)(T_1(x)-x)]\mathrm{d}x,$$
$$\widehat{\rho}_-^1 = \frac{1}{n-1}\sum_{i=2}^{n}\int_\Omega (T_i(x)-x)(x-T_{i-1}^{-1}(x))\mathrm{d}x, \quad \rho_-^1 = \int_\Omega \mathbb{E}[(T_2(x)-x)(x-T_1^{-1}(x))]\mathrm{d}x.$$

Then, $\widehat{\alpha}_n$ can be written as

$$\widehat{\alpha}_n = \widehat{\alpha}_+ \mathbb{I}\{l_+(\widehat{\alpha}_+) \le l_-(\widehat{\alpha}_-)\} + \widehat{\alpha}_- \mathbb{I}\{l_+(\widehat{\alpha}_+) > l_-(\widehat{\alpha}_-)\},$$

where $\widehat{\alpha}_+ = \widehat{\rho}^1_+/\widehat{\rho}^0_+$, $\widehat{\alpha}_- = \widehat{\rho}^1_-/\widehat{\rho}^0_-$, and $l_+(\cdot)$ and $l_-(\cdot)$ are respectively defined by

$$l_+(\alpha) = \sum_{i=2}^{n} \int_\Omega \left(T_i(x) - x - \alpha(T_{i-1}(x) - x)\right)^2 \mathrm{d}x,$$

and

$$l_-(\alpha) = \sum_{i=2}^{n} \int_\Omega \left(T_i(x) - x - \alpha(x - T_{i-1}^{-1}(x))\right)^2 \mathrm{d}x.$$

The proof consists of the following three steps:

1. for any $\widehat{\rho} \in \{\widehat{\rho}^0_+, \widehat{\rho}^0_-, \widehat{\rho}^1_+, \widehat{\rho}^1_-\}$, and corresponding true value $\rho \in \{\rho^0_+, \rho^0_-, \rho^1_+, \rho^1_-\}$, we have that for $\forall \lambda > 0$, there exist some universal constants $c_1, c_2 > 0$ such that
$$\mathbb{P}(|\widehat{\rho} - \rho| \ge \lambda) \le c_1 \exp\left(-(n-1)\lambda^2 c_2\right); \tag{S.1}$$
2. the probability of misclassification is negligible, i.e. for sufficiently large n,
$$\mathbb{P}(\widehat{\alpha}_n = \widehat{\alpha}_+, \alpha_0 \le 0) + \mathbb{P}(\widehat{\alpha}_n = \widehat{\alpha}_-, \alpha_0 > 0) = o_p(n^{-1/2});$$
3. the asymptotic normality of $\widehat{\alpha}_+$ when $\alpha_0 > 0$, and that of $\widehat{\alpha}_-$ when $\alpha_0 \le 0$.

Step 1. Without loss of generality, we only prove for the case of $\widehat{\rho}^0_+$ as others can be verified similarly. In particular, we aim to show: for any $\lambda > 0$, there exist two constants $c_1', c_2' > 0$ such that

$$\mathbb{P}\left(\left|\sum_{i=2}^{n} \int_\Omega (T_i(x) - x)^2 \mathrm{d}x - (n-1)\rho^0_+\right| \ge (n-1)\lambda\right) \le c_1' \exp\left(-(n-1)\lambda^2 c_2'\right). \tag{S.2}$$

Note that $\mathbb{E}\left(\int_\Omega (T_i(x) - x)^2 \mathrm{d}x\right) - \rho^0_+ = 0$, and that given $T_1 = t_1$, $T_i = \widetilde{\phi}_{i,i-1}(t_1)$. Then, by Proposition 1 in Wu and Shao (2004), it suffices to show that

$$\sup_{t_1 \in \mathcal{T}} \sum_{i=2}^{\infty} \left|\mathbb{E}\left[\int_\Omega (\widetilde{\phi}_{i,i-1}(t_1) - x)^2 \mathrm{d}x - \rho^0_+\right]\right| < \infty. \tag{S.3}$$

Define $h(t) = \mathbb{E}\left[\int_\Omega (T_2(x) - x)^2 \mathrm{d}x \middle| T_1 = t\right]$. Recall $\mathbb{E}\int_\Omega (T_{\epsilon_i}(x) - x)^2 \mathrm{d}x = \sigma_1^2 < \infty$. Then, we have

$$h(t_1) = \mathbb{E}\left\{\int_\Omega \left[T_{\epsilon_2}(x + \alpha_0(T_1(x) - x)) - x\right]^2 \mathrm{d}x \middle| T_1 = t_1\right\}$$

$$
\begin{aligned}
=&\mathbb{E}\left\{\int_\Omega \left[T_{\epsilon_2}(x+\alpha_0(T_1(x)-x))\right]^2 \mathrm{d}x \middle| T_1=t_1\right\}\\
+&\mathbb{E}\left[\int_\Omega x^2 \mathrm{d}x \middle| T_1=t_1\right]-2\mathbb{E}\left[\int_\Omega xT_{\epsilon_2}(x+\alpha_0(T_1(x)-x))\mathrm{d}x \middle| T_1=t_1\right]\\
=&\mathbb{E}\left\{\int_\Omega \left[T_{\epsilon_2}(x+\alpha_0(t_1(x)-x))\right]^2 \mathrm{d}x\right\}\\
+&\int_\Omega x^2\mathrm{d}x-2\left[\int_\Omega x(x+\alpha_0(t_1(x)-x))\mathrm{d}x\right]\\
=&\mathbb{E}\left\{\int_\Omega \left[T_{\epsilon_2}(x+\alpha_0(t_1(x)-x))\right]^2 \mathrm{d}x\right\}-\int_\Omega x^2\mathrm{d}x-2\alpha_0\int_\Omega x(t_1(x)-x)\mathrm{d}x,
\end{aligned}
$$

where the third equality holds by noting that T_{ϵ_2} is independent of t_1, and that $\mathbb{E}[T_{\epsilon_2}(x)]=x$ for any $x\in\Omega$.

Therefore, for $t_1\neq t_1'\in\mathcal{T}$, we obtain that

$$
\begin{aligned}
h(t_1)-h(t_1')=&\mathbb{E}\left\{\int_\Omega \left[T_{\epsilon_2}(x+\alpha_0(t_1(x)-x))\right]^2 \mathrm{d}x\right\}-\mathbb{E}\left\{\int_\Omega \left[T_{\epsilon_2}(x+\alpha_0(t_1'(x)-x))\right]^2 \mathrm{d}x\right\}\\
&-2\alpha_0\int_\Omega x[t_1(x)-t_1'(x)]\mathrm{d}x\\
:=&G_1+G_2.
\end{aligned}
$$

For the term G_1, we have

$$
\begin{aligned}
G_1=&\Big|\mathbb{E}\left\{\int_\Omega \left[T_{\epsilon_2}(x+\alpha_0(t_1(x)-x))-T_{\epsilon_2}(x+\alpha_0(t_1'(x)-x))\right]\right.\\
&\left.\times\left[T_{\epsilon_2}(x+\alpha_0(t_1(x)-x))+T_{\epsilon_2}(x+\alpha_0(t_1'(x)-x))\right]\mathrm{d}x\right\}\Big|\\
\leq&\mathbb{E}\left\{\int_\Omega \left|T_{\epsilon_2}(x+\alpha_0(t_1(x)-x))-T_{\epsilon_2}(x+\alpha_0(t_1'(x)-x))\right|\right.\\
&\left.\times\left|T_{\epsilon_2}(x+\alpha_0(t_1(x)-x))+T_{\epsilon_2}(x+\alpha_0(t_1'(x)-x))\right|\mathrm{d}x\right\}\\
\leq&2\overline{B}\mathbb{E}\left[\int_\Omega \left|(T_{\epsilon_2}(x+\alpha_0(t_1(x)-x))-T_{\epsilon_2}(x+\alpha_0(t_1'(x)-x)))\right|\mathrm{d}x\right]\\
=&2\overline{B}\mathbb{E}[d_1(\phi_{\epsilon_2}(t_1),\phi_{\epsilon_2}(t_1'))]\\
\leq&2\overline{B}Cr d_1(t_1,t_1'),
\end{aligned}
$$

where the second inequality holds by the boundedness of the maps in $\mathcal{T}$, i.e. for any $t\in\mathcal{T}$, there exists a constant $\overline{B}>0$ such that $\sup_{x\in\Omega}|t(x)|\leq\overline{B}$, the third inequality holds by Assumption 3.1.

By the compactness of Ω, we can easily see that $G_2\leq 2\alpha_0\overline{B}d_1(t_1,t_1')$. Hence, we obtain

$$
|h(t_1)-h(t_1')|\leq 2\overline{B}(|\alpha_0|+Cr)d_1(t_1,t_1').
$$

By an induction argument, we can show that

$$
|h(\widetilde{\phi}_{i,i-1}(t_1))-h(\widetilde{\phi}_{i,i-1}(t_1'))|\leq 2\overline{B}(|\alpha_0|+Cr)d_1(\widetilde{\phi}_{i-1,i-2}(t_1),\widetilde{\phi}_{i-1,i-2}(t_1')). \tag{S.4}
$$

Let $\mathcal{F}_i$ be the sigma-field generated by $\{T_{\epsilon_j}, 2 \le j \le i\}$, and note that given T_1, $\mathcal{F}_i$ uniquely determines T_i. Therefore, for $i \ge 3$, using the tower property of conditional expectation, we have

$$\begin{aligned}
&\left|\mathbb{E}\left[\int_\Omega (\widetilde{\phi}_{i,i-1}(t_1) - x)^2 \mathrm{d}x - \rho_+^0\right]\right| \\
=&\left|\mathbb{E}_{T_1'}\left\{\mathbb{E}\left[\int_\Omega (\widetilde{\phi}_{i,i-1}(t_1) - x)^2 \mathrm{d}x - \int_\Omega (\widetilde{\phi}_{i,i-1}(T_1') - x)^2 \mathrm{d}x\right]\right\}\right| \\
=&\left|\mathbb{E}_{T_1'}\mathbb{E}\left\{\mathbb{E}_{\epsilon_i}\left[\int_\Omega (\widetilde{\phi}_{i,i-1}(t_1) - x)^2 \mathrm{d}x - \int_\Omega (\widetilde{\phi}_{i,i-1}(T_1') - x)^2 \mathrm{d}x \middle| \mathcal{F}_{i-1}\right]\right\}\right| \\
\le& 2\overline{B}(\alpha_0 + Cr)\mathbb{E}_{T_1'}\left\{\mathbb{E} d_1(\phi_{i-1,i-2}(t_1), \phi_{i-1,i-2}(T_1'))\right\} \\
\le& 2\overline{B}(\alpha_0 + Cr)Cr^{i-2}\mathbb{E}_{T_1'} d_1(t_1, T_1'),
\end{aligned}$$

where the first inequality holds by (S.4), and the last by Assumption 3.1.

Therefore, by the boundedness of $\mathbb{E}_{T_1'} d_1(t_1, T_1')$, and that $r \in (0,1)$ by Assumption 3.1, (S.3) thus follows, which further implies (S.2).

Hence,

$$\mathbb{P}(|\widehat{\rho}_+^0 - \rho_+^0| \ge \lambda) \le c_1' \exp\left(-(n-1)\lambda^2 c_2'\right).$$

Similarly, for $\widehat{\rho}_-^0, \widehat{\rho}_+^1, \widehat{\rho}_-^1$, we can show that

$$\begin{aligned}
\mathbb{P}(|\widehat{\rho}_-^0 - \rho_-^0| \ge \lambda) &\le c_3' \exp\left(-(n-1)\lambda^2 c_4'\right), \\
\mathbb{P}(|\widehat{\rho}_+^1 - \rho_+^1| \ge \lambda) &\le c_5' \exp\left(-(n-1)\lambda^2 c_6'\right), \\
\mathbb{P}(|\widehat{\rho}_-^1 - \rho_-^1| \ge \lambda) &\le c_7' \exp\left(-(n-1)\lambda^2 c_8'\right),
\end{aligned}$$

where $c_3', \cdots, c_8'$ are positive constants.

To unify, let $c_1 = \max\{c_1', c_3', c_5', c_7'\}$ and $c_2 = \min\{c_2', c_4', c_6', c_8'\}$. Then, for any corresponding pair

$$\widehat{\rho} \in \{\widehat{\rho}_+^0, \widehat{\rho}_-^0, \widehat{\rho}_+^1, \widehat{\rho}_-^1\}, \qquad \rho \in \{\rho_+^0, \rho_-^0, \rho_+^1, \rho_-^1\},$$

we have, for any $\lambda > 0$,

$$\mathbb{P}(|\widehat{\rho} - \rho| \ge \lambda) \le c_1 \exp\left(-(n-1)\lambda^2 c_2\right). \tag{S.5}$$

Step 2. We next show that the probability of misclassification is negligible. Without loss of generality, we only prove that $\mathbb{P}(\widehat{\alpha}_n = \widehat{\alpha}_-, \alpha_0 > 0) = o_p(n^{-1/2})$.

Note that the estimator $\widehat{\alpha}_n = \widehat{\alpha}_-$ if $l_+(\widehat{\alpha}_+) \ge l_-(\widehat{\alpha}_-)$, where we have

$$l_+(\widehat{\alpha}_+) = \frac{1}{n-1}\sum_{i=2}^n \int_\Omega (T_i(x) - x)^2 \mathrm{d}x - \frac{(\widehat{\rho}_+^1)^2}{\widehat{\rho}_+^0}, \tag{S.6}$$

$$l_-(\widehat{\alpha}_-) = \frac{1}{n-1}\sum_{i=2}^n \int_\Omega (T_i(x) - x)^2 \mathrm{d}x - \frac{(\widehat{\rho}_-^1)^2}{\widehat{\rho}_-^0}. \tag{S.7}$$

It's clear that $l_+(\widehat{\alpha}_+) \ge l_-(\widehat{\alpha}_-)$ if and only if $(\widehat{\rho}_+^1)^2/\widehat{\rho}_+^0 \le (\widehat{\rho}_-^1)^2/\widehat{\rho}_-^0$, which is equivalent to $(\widehat{\rho}_+^1)^2\widehat{\rho}_-^0 \le (\widehat{\rho}_-^1)^2\widehat{\rho}_+^0$.

Note by the proof of Theorem 2 in Zhu and Müller (2023), if $\mathbb{E}\int_\Omega (T_i(x)-x)^2 \mathrm{d}x \geq \underline{c}$, then there exists a constant $c'>0$ such that $(\rho_+^1)^2\rho_-^0 - (\rho_-^1)^2\rho_+^0 \geq c'$. We have that

$$\begin{aligned}
\mathbb{P}\left((\widehat{\rho}_+^1)^2\widehat{\rho}_-^0 - (\widehat{\rho}_-^1)^2\widehat{\rho}_+^0 \leq 0\right) \leq & \mathbb{P}\left((\rho_+^1)^2\rho_-^0 - (\rho_-^1)^2\rho_+^0 - (\widehat{\rho}_+^1)^2\widehat{\rho}_-^0 + (\widehat{\rho}_-^1)^2\widehat{\rho}_+^0 \geq c'\right) \\
\leq & \mathbb{P}\left(|(\rho_+^1)^2\rho_-^0 - (\widehat{\rho}_+^1)^2\widehat{\rho}_-^0| \geq \frac{c'}{2}\right) + \mathbb{P}\left(|(\rho_-^1)^2\rho_+^0 - (\widehat{\rho}_-^1)^2\widehat{\rho}_+^0| \geq \frac{c'}{2}\right) \\
\leq & \mathbb{P}\left(|(\rho_+^1)^2\rho_-^0 - (\rho_+^1)^2\widehat{\rho}_-^0| \geq \frac{c'}{6}\right) + \mathbb{P}\left(|(\rho_+^1)^2\widehat{\rho}_-^0 - \rho_+^1\widehat{\rho}_+^1\widehat{\rho}_-^0| \geq \frac{c'}{6}\right) \\
& + \mathbb{P}\left(|\rho_+^1\widehat{\rho}_+^1\widehat{\rho}_-^0 - (\widehat{\rho}_+^1)^2\widehat{\rho}_-^0| \geq \frac{c'}{6}\right) + \mathbb{P}\left(|(\rho_-^1)^2\rho_+^0 - (\rho_-^1)^2\widehat{\rho}_+^0| \geq \frac{c'}{6}\right) \\
& + \mathbb{P}\left(|(\rho_-^1)^2\widehat{\rho}_+^0 - \rho_-^1\widehat{\rho}_-^1\widehat{\rho}_+^0| \geq \frac{c'}{6}\right) + \mathbb{P}\left(|\rho_-^1\widehat{\rho}_-^1\widehat{\rho}_+^0 - (\widehat{\rho}_-^1)^2\widehat{\rho}_+^0| \geq \frac{c'}{6}\right) \\
:= & \sum_{i=1}^6 P_i.
\end{aligned}$$

We now provide upper bound for each term in the above inequality. By Cauchy-Schwarz inequality, we can show that for some constant $\bar{c}>0$, $\max\{|\rho_+^1|, |\rho_+^0|, |\rho_-^1|, |\rho_-^0|\} \leq \max\{\rho_+^0, \rho_-^0\} \leq \bar{c}$. Hence,

$$P_1 = \mathbb{P}(|(\rho_+^1)^2\rho_-^0 - (\rho_+^1)^2\widehat{\rho}_-^0| \geq \frac{c'}{6}) \leq \mathbb{P}(|\rho_-^0 - \widehat{\rho}_-^0| \geq \frac{c'}{6\bar{c}^2}) \leq c_1 \exp\big(-\frac{(n-1)c_2(c')^2}{36\bar{c}^4}\big),$$

where the first inequality holds since $|\rho_+^1| \leq \bar{c}$, and the second holds by (S.5) with $\lambda = c'/(6\bar{c}^2)$.

For P_2, we have

$$\begin{aligned}
P_2 = \mathbb{P}(|(\rho_+^1)^2\widehat{\rho}_-^0 - \rho_+^1\widehat{\rho}_+^1\widehat{\rho}_-^0| \geq c'/6) \leq & \mathbb{P}(|\widehat{\rho}_-^0| \geq 2\bar{c}) + \mathbb{P}(|\widehat{\rho}_-^0||(\rho_+^1)^2 - \rho_+^1\widehat{\rho}_+^1| \geq c'/6, |\widehat{\rho}_-^0| \leq 2\bar{c}) \\
\leq & \mathbb{P}(|\widehat{\rho}_-^0 - \rho_-^0| \geq \bar{c}) + \mathbb{P}(|\rho_+^1 - \widehat{\rho}_+^1| \geq \frac{c'}{12\bar{c}^2}) \\
\leq & c_1 \exp\big(-\bar{c}^2(n-1)c_2\big) + c_1 \exp\big(-\frac{(c')^2(n-1)c_2}{144\bar{c}^4}\big),
\end{aligned}$$

where the second inequality holds since $|\rho_+^1|$ and $|\rho_-^0|$ are upper bounded by $\bar{c}$, and the third holds by (S.5). Next,

$$\begin{aligned}
P_3 = & \mathbb{P}\{|\rho_+^1\widehat{\rho}_+^1\widehat{\rho}_-^0 - (\widehat{\rho}_+^1)^2\widehat{\rho}_-^0| \geq c'/6\} \\
\leq & \mathbb{P}(|\widehat{\rho}_-^0| \geq 2\bar{c}) + \mathbb{P}(|\widehat{\rho}_+^1| \geq 2\bar{c}) + \mathbb{P}\left(|\rho_+^1 - \widehat{\rho}_+^1| \geq \frac{c'}{24\bar{c}^2}\right) \\
\leq & \mathbb{P}(|\widehat{\rho}_-^0 - \rho_-^0| \geq \bar{c}) + \mathbb{P}(|\widehat{\rho}_+^1 - \rho_+^1| \geq \bar{c}) + \mathbb{P}(|\rho_+^1 - \widehat{\rho}_+^1| \geq \frac{c'}{24\bar{c}^2}) \\
\leq & c_1 \exp\big(-\bar{c}^2(n-1)c_2\big) + c_1 \exp\big(-\bar{c}^2(n-1)c_2\big) \\
& + c_1 \exp\big(-\frac{(c')^2(n-1)c_2}{576\bar{c}^4}\big).
\end{aligned}$$

The proof for the terms P_4, P_5, P_6 are similar.

Therefore, by setting $C_1 = c_1$, and $C_2 = \min\{\bar{c}^2 c_2, (c')^2 c_2/576\bar{c}^4, (c')^2 c_2/36\bar{c}^4\}$, then

$$P_i \leq 3C_1 \exp\left(-(n-1)C_2\right), \quad i = 1, 2, \cdots, 6.$$

Therefore, we have that

$$\mathbb{P}(\widehat{\alpha}_n = \widehat{\alpha}_-, \alpha_0 > 0) \leq \mathbb{P}((\widehat{\rho}_+^1)^2 \widehat{\rho}_-^0 - (\widehat{\rho}_-^1)^2 \widehat{\rho}_+^0 \leq 0) \leq 18C_1 \exp\left(-C_2(n-1)\right).$$

Hence, if $\alpha_0 > 0$, we have

$$\mathbb{P}(\widehat{\alpha}_n \neq \widehat{\alpha}_+) \leq 18C_1 \exp\left(-C_2(n-1)\right) = o(n^{-1/2}), \tag{S.8}$$

where the second equality holds for sufficiently large n.

Step 3. Considering the case when $\alpha_0 > 0$, we have

$$\begin{aligned}\sqrt{n}(\widehat{\alpha}_n - \alpha_0) =& \sqrt{n}(\widehat{\alpha}_+ \mathbb{I}\{\widehat{\alpha}_n = \widehat{\alpha}_+\} + \widehat{\alpha}_- \mathbb{I}\{\widehat{\alpha}_n \neq \widehat{\alpha}_+\} - \alpha_0) \\ =& \sqrt{n}(\widehat{\alpha}_+ - \alpha_0) + \sqrt{n}(\widehat{\alpha}_+ \mathbb{I}\{\widehat{\alpha}_n \neq \widehat{\alpha}_+\} + \widehat{\alpha}_- \mathbb{I}\{\widehat{\alpha}_n \neq \widehat{\alpha}_+\}).\end{aligned}$$

Since the Cauchy-Schwarz inequality ensures that $|\widehat{\alpha}_+|$ and $|\widehat{\alpha}_-|$ are upper bounded by 1, we have that

$$\sqrt{n}(\widehat{\alpha}_+ \mathbb{I}\{\widehat{\alpha}_n \neq \widehat{\alpha}_+\} + \widehat{\alpha}_- \mathbb{I}\{\widehat{\alpha}_n \neq \widehat{\alpha}_+\}) \leq 2\sqrt{n}\mathbb{I}\{\widehat{\alpha}_n \neq \widehat{\alpha}_+\},$$

which is $o_p(1)$ by Markov inequality in view of (S.8).

Therefore,

$$\begin{aligned}&\sqrt{n}(\widehat{\alpha}_n - \alpha_0) = \sqrt{n}(\widehat{\alpha}_+ - \alpha_0) + o_p(1) \\ =& \left\{\frac{1}{n}\sum_{i=1}^{n}\int_\Omega (T_i(x) - x)^2 \mathrm{d}x\right\}^{-1} \left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n-1}\int_\Omega \{T_{i+1}(x) - [\alpha_0 \odot T_i](x)\}(T_i(x) - x)\mathrm{d}x\right\} + o_p(1) \\ =& \Xi^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n-1}\int_\Omega \{T_{i+1}(x) - [\alpha_0 \odot T_i](x)\}(T_i(x) - x)\mathrm{d}x + o_p(1).\end{aligned}$$

Furthermore, we have that $\{\int_\Omega \{T_{i+1}(x) - [\alpha_0 \odot T_i](x)\}(T_i(x) - x)]\mathrm{d}x\}$ forms a M.D.S.. To see this,

$$\begin{aligned}&\mathbb{E}\left\{\int_\Omega \{T_{i+1}(x) - [\alpha_0 \odot T_{i-1}](x)\}(T_i(x) - x)]\mathrm{d}x \middle| \mathcal{F}_i\right\} \\ =& \int_\Omega \mathbb{E}\left\{\{T_{i+1}(x) - [\alpha_0 \odot T_{i-1}](x)\}(T_i(x) - x)]\mathrm{d}x \middle| \mathcal{F}_i\right\} \\ =& \int_\Omega \{\mathbb{E}\left[T_{i+1}(x) \middle| \mathcal{F}_i\right] - [\alpha_0 \odot T_{i-1}](x)\}(T_i(x) - x)]\mathrm{d}x = 0.\end{aligned}$$

For $\alpha_0 \leq 0$, similarly, we can show that

$$\sqrt{n}(\widehat{\alpha}_n - \alpha_0) = \sqrt{n}(\widehat{\alpha}_+ \mathbb{I}_{\{\widehat{\alpha} \neq \widehat{\alpha}_-\}} + \widehat{\alpha}_- \mathbb{I}_{\{\widehat{\alpha} = \widehat{\alpha}_-\}} - \alpha_0)$$

$$
\begin{aligned}
&= \sqrt{n}(\widehat{\alpha}_- - \alpha_0) + o_p(1) \\
&= \left\{ \frac{1}{n} \sum_{i=1}^{n} \int_\Omega (x - T_i^{-1}(x))^2 \mathrm{d}x \right\}^{-1} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n-1} \int_\Omega \{T_{i+1}(x) - [\alpha \odot T_i](x)\}(x - T_i^{-1}(x))] \mathrm{d}x \right\} + o_p(1).
\end{aligned}
$$

In addition, we have that

$$
\begin{aligned}
&\mathbb{E}\left\{ \int_\Omega \{T_{i+1}(x) - [\alpha_0 \odot T_i](x)\}(x - T_i^{-1}(x)) \mathrm{d}x | \mathcal{F}_i \right\} \\
&= \int_\Omega \left\{ \{\mathbb{E}[T_{i+1}(x)|\mathcal{F}_i] - [\alpha_0 \odot T_i](x)\}(x - T_i^{-1}(x)) \mathrm{d}x \right\} = 0.
\end{aligned}
$$

□

S.2 Proof of Theorem 4.1 and Theorem 4.2

Note that Theorem 4.1 is a special case Theorem 4.2 by setting $f_n = l_n = n$, hence we only prove for Theorem 4.2.

Recall that we estimate α_0 based on the first f_n observations, and use the last l_n residuals to obtain the sample ACF. Define

$$
\widehat{\gamma}_{l_n}(k) = \frac{1}{l_n - k} \sum_{i=n-l_n+1}^{n-k} \int_\Omega \left[\left(\widehat{T}_{\epsilon_{i+k}}^{f_n}(x) - x \right) \left(\widehat{T}_{\epsilon_i}^{f_n}(x) - x \right) \right] \mathrm{d}x,
$$

and

$$
\widetilde{\gamma}_{l_n}(k) = \frac{1}{l_n - k} \sum_{i=n-l_n+1}^{n-k} \int_\Omega \left[\left(T_{\epsilon_{i+k}}(x) - x \right) \left(T_{\epsilon_i}(x) - x \right) \right] \mathrm{d}x.
$$

Clearly, we have $T_{\epsilon_i}(x) = G_i(\alpha_0, x)$ and $\widehat{T}_{\epsilon_i}^{f_n} = G_i(\widehat{\alpha}_{f_n}, x) = T_i \circ [\widehat{\alpha}_{f_n} \odot T_{i-1}]^{-1}$. Note that by Theorem 3.1, we have we have that as $f_n \to \infty$,

$$
\sqrt{f_n}(\widehat{\alpha}_{f_n} - \alpha) = \frac{1}{\sqrt{f_n}} \sum_{i=1}^{f_n} m_i(\alpha_0) + o_p(1). \tag{S.9}
$$

Then, we decompose

$$\begin{aligned}\sqrt{l_n}\widehat{\gamma}_{l_n}(k) =&\frac{1}{\sqrt{l_n}}\sum_{i=n-l_n+1}^{n-k}\int_\Omega\left[\left(\widehat{T}^{f_n}_{\epsilon_{i+k}}(x)-T_{\epsilon_{i+k}}(x)\right)\left(T_{\epsilon_i}(x)-x\right)\right]\mathrm{d}x\\ &+\frac{1}{\sqrt{l_n}}\sum_{i=n-l_n+1}^{n-k}\int_\Omega\left[\left(T_{\epsilon_{i+k}}(x)-x\right)\left(\widehat{T}^{f_n}_{\epsilon_i}(x)-T_{\epsilon_i}(x)\right)\right]\mathrm{d}x\\ &+\frac{1}{\sqrt{l_n}}\sum_{i=n-l_n+1}^{n-k}\int_\Omega\left[\left(\widehat{T}^{f_n}_{\epsilon_{i+k}}(x)-T_{\epsilon_{i+k}}(x)\right)\left(\widehat{T}^{f_n}_{\epsilon_i}(x)-T_{\epsilon_i}(x)\right)\right]\mathrm{d}x\\ &+\frac{1}{\sqrt{l_n}}\sum_{i=n-l_n+1}^{n-k}\int_\Omega\left[\left(T_{\epsilon_{i+k}}(x)-x\right)\left(T_{\epsilon_i}(x)-x\right)\right]\mathrm{d}x\\ =:&Z_1+Z_2+Z_3+\sqrt{l_n}\widetilde{\gamma}_{l_n}(k).\end{aligned}\tag{S.10}$$

(1) For Z_1, by Taylor expansion, we obtain that

$$\begin{aligned}Z_1=&\frac{1}{\sqrt{l_n}}\sum_{i=n-l_n+1}^{n-k}\int_\Omega\left[\left(\widehat{T}_{\epsilon_{i+k}}(x)-T_{\epsilon_{i+k}}(x)\right)\left(T_{\epsilon_i}(x)-x\right)\right]\mathrm{d}x\\ =&\sqrt{l_n}(\widehat{\alpha}_{f_n}-\alpha_0)\cdot\frac{1}{l_n}\sum_{i=n-l_n+1}^{n-k}\int_\Omega\left[\left(T_{\epsilon_i}(x)-x\right)g_{i+k}(\alpha^*,x)\right]\mathrm{d}x,\end{aligned}$$

where $\alpha^*=\alpha_0+\eta(\widehat{\alpha}_{f_n}-\alpha_0)$ for some $\eta\in[0,1]$.

We then show that

$$\frac{1}{l_n}\sum_{i=n-l_n+1}^{n-k}\int_\Omega\left[\left(T_{\epsilon_i}(x)-x\right)|g_{i+k}(\alpha^*,x)-g_{i+k}(\alpha_0,x)|\right]\mathrm{d}x=o_p(1).\tag{S.11}$$

In particular, since $|\widehat{\alpha}_{f_n}-\alpha_0|=o_p(1)$, it suffices to show

$$\limsup_{\delta\downarrow0}\limsup_{n\to\infty}\frac{1}{l_n}\sum_{i=n-l_n+1}^{n-k}\mathbb{E}\left[\int_\Omega(T_{\epsilon_i}(x)-x)(g_{i+k}(\alpha^*,x)-g_{i+k}(\alpha_0,x))\mathbb{I}\{|\widehat{\alpha}_{f_n}-\alpha_0|\le\delta\}\mathrm{d}x\right]=0.$$

Note for any $\delta>0$, by Cauchy-Schwarz inequality, we have

$$\begin{aligned}&\limsup_{n\to\infty}\frac{1}{l_n}\sum_{i=n-l_n+1}^{n-k}\mathbb{E}\left[\int_\Omega(T_{\epsilon_i}(x)-x)(g_{i+k}(\alpha^*,x)-g_{j+k}(\alpha_0,x))\mathbb{I}\{|\widehat{\alpha}_{f_n}-\alpha_0|\le\delta\}\mathrm{d}x\right]\\ \le&\limsup_{n\to\infty}\frac{\sigma_1}{l_n}\sum_{i=n-l_n+1}^{n-k}\left(\mathbb{E}\int_\Omega[g_{i+k}(\alpha^*,x)-g_{i+k}(\alpha_0,x)]^2\mathbb{I}\{|\widehat{\alpha}_{f_n}-\alpha_0|\le\delta\}\mathrm{d}x\right)^{1/2}\\ \le&\sigma_1\left(\mathbb{E}\sup_{|\alpha_1-\alpha|\le\delta}\int_\Omega[g_i(\alpha_1,x)-g_i(\alpha_0,x)]^2\mathrm{d}x\right)^{1/2},\end{aligned}$$

which goes to zero as $\delta \downarrow 0$ under the assumption that $\mathbb{E}\sup_{|\alpha_1-\alpha_0|\leq\delta_0}\int_\Omega g_i^2(\alpha_1,x)\mathrm{d}x<\infty$.

Hence, with (S.9) and (S.11), by the ergodic theorem, we can write

$$\begin{aligned} Z_1 =&\sqrt{l_n}(\widehat{\alpha}_{f_n}-\alpha_0)\cdot\left(\frac{1}{l_n}\sum_{i=n-l_n+1}^{n-k}\int_\Omega\left[(T_{\epsilon_i}(x)-x)\,g_{i+k}(\alpha_0,x)\right]\mathrm{d}x+o_p(1)\right)\\ =&\frac{\sqrt{l_n}}{\sqrt{f_n}}\sum_{i=1}^{f_n}\frac{1}{\sqrt{f_n}}\mathbb{E}\left[\int_\Omega(T_{\epsilon_i}(x)-x)\,g_{i+k}(\alpha_0,x)\mathrm{d}x\right]m_i(\alpha_0)+o_p(1+\sqrt{l_n/f_n}). \end{aligned} \tag{S.12}$$

(2) By similar arguments, we can obtain that

$$\begin{aligned} Z_2 =&\frac{\sqrt{l_n}}{\sqrt{f_n}}\sum_{i=1}^{f_n}\frac{1}{\sqrt{f_n}}\mathbb{E}\left[\int_\Omega\left(T_{\epsilon_{i+k}}(x)-x\right)g_i(\alpha_0,x)\mathrm{d}x\right]m_i(\alpha_0)+o_p(1+\sqrt{l_n/f_n})\\ =&o_p(1+\sqrt{l_n/f_n}), \end{aligned} \tag{S.13}$$

where the second equality holds by noting that $T_{\epsilon_{i+k}}(x)$ is independent of $g_i(\alpha;x)\in\mathcal{F}_i$ for $k>0$.

(3) Applying the similar arguments, we have that

$$Z_3=\frac{\sqrt{l_n}}{f_n}\left(\sqrt{f_n}(\widehat{\alpha}_{f_n}-\alpha)\right)^2\frac{1}{l_n}\sum_{i=n-l_n+1}^{n-k}\int_\Omega\left[g_{i+k}(\alpha_1^*,x)g_i(\alpha_2^*,x)\right]\mathrm{d}x+o_p(1)=O_p(\sqrt{l_n}/f_n)+o_p(1), \tag{S.14}$$

where $\alpha_l^*=\alpha_0+\eta_l(\widehat{\alpha}-\alpha_0)$ for some $\eta_l\in[0,1], l=1,2$.

Summarizing (S.10)-(S.14), and under the condition that $l_n/f_n<\infty$, we obtain that,

$$\begin{aligned} \sqrt{l_n}\widehat{\gamma}_{l_n}(k)=&\frac{1}{\sqrt{l_n}}\sum_{i=n-l_n+1}^{n-k}\int_\Omega\left[\left(\widehat{T}_{\epsilon_{i+k}}(x)-x\right)\left(\widehat{T}_{\epsilon_i}(x)-x\right)\right]\mathrm{d}x\\ =&\mu_k\frac{\sqrt{l_n}}{f_n}\sum_{i=1}^{f_n}m_i(\alpha_0)+\sqrt{l_n}\widetilde{\gamma}_{l_n}(k)+o_p(1). \end{aligned}$$

By similar arguments as the proof of Theorem 1 in Davis and Fernandes (2025), we obtain that

$$\sqrt{l_n}\left(\widetilde{\gamma}_{l_n}(1),\cdots,\widetilde{\gamma}_{l_n}(K),\frac{1}{f_n}\sum_{i=1}^{f_n}m_i(\alpha)\right)\to_d\mathcal{N}\left(\mathbf{0},\begin{pmatrix}\sigma_2^4I_K & k_{ov}M_2\\ k_{ov}M_2^T & k_{ra}\mathbb{E}[m_i^2(\alpha_0)]\end{pmatrix}\right).$$

Finally, we note that $\widehat{\rho}_{l_n}(k)=\widehat{\gamma}_{l_n}(k)/\widehat{\gamma}_{l_n}(0)\cdot l_n/(l_n-k)$, where it is easy to show that

$$\widehat{\gamma}_{l_n}(0)=\frac{1}{l_n}\sum_{i=n-l_n+1}^{n}\int_\Omega\left[\left(\widehat{T}_{\epsilon_i}(x)-x\right)^2\right]\mathrm{d}x\to_p\int_\Omega\mathbb{E}\left[(T_{\epsilon_i}(x)-x)^2\right]\mathrm{d}x=\sigma_1^2.$$

Therefore, the result follows by the continuous mapping theorem.

S.3 Additional Discussion on ATM(2) Model

A natural second-order autoregressive transport map model is the ATM(2) model

$$T_i = T_{\epsilon_i} \circ [\alpha_1 \odot T_{i-1}] \circ [\alpha_2 \odot T_{i-2}], \qquad i = 3, \ldots, n,$$

with parameter vector $\theta = (\alpha_1, \alpha_2)^\top$. The least-squares estimator is the $p = 2$ version of the criterion in Zhu and Müller (2023) and can be defined by

$$\widehat{\theta}_n \in \arg\min_{\alpha_1, \alpha_2} l_2(\alpha_1, \alpha_2), \qquad l_2(\alpha_1, \alpha_2) = \sum_{i=3}^{n} \int_\Omega \{T_i(x) - [\alpha_1 \odot T_{i-1}] \circ [\alpha_2 \odot T_{i-2}](x)\}^2 \, dx.$$

Given $\widehat{\theta}_n = (\widehat{\alpha}_{1,n}, \widehat{\alpha}_{2,n})^\top$, the corresponding residual map is

$$\widehat{T}^{(2)}_{\epsilon_i} = T_i \circ [\widehat{\alpha}_{2,n} \odot T_{i-2}]^{-1} \circ [\widehat{\alpha}_{1,n} \odot T_{i-1}]^{-1}.$$

These formulas show that the residual-ACF diagnostic can be implemented for ATM(2). A natural split-sample analogue of the ATM(1) statistic is

$$Q_n^{S,(2)} = l_n \, \widehat{\rho}^{(2)}_{l_n}(\widehat{\theta}_{f_n})^\top \{\widehat{\Sigma}_K^{S,(2)}\}^{-1} \widehat{\rho}^{(2)}_{l_n}(\widehat{\theta}_{f_n}), \tag{S.15}$$

where $\widehat{\rho}^{(2)}_{l_n}(\widehat{\theta}_{f_n})$ is the vector of ATM(2) residual autocorrelations computed from the second split using the first-split estimator $\widehat{\theta}_{f_n}$. However, the sample-splitting test used for ATM(1) does not extend by a direct substitution.

The main difficulty can be seen directly from the first-order expansion. For simplicity, consider the sign region $\alpha_{1,0} > 0$ and $\alpha_{2,0} > 0$; the other sign configurations are analogous, with $T(x) - x$ replaced by the corresponding inverse-map displacement in the contraction operator.

Let

$$C_a(T)(x) = x + a\{T(x) - x\}, \qquad F_i(\theta, x) = C_{\alpha_1}(T_i) \circ C_{\alpha_2}(T_{i-1})(x),$$

where $\theta = (\alpha_1, \alpha_2)^\top$. The first-split ATM(2) least-squares estimator satisfies the estimating equation

$$0 = \sum_{i=2}^{f_n - 1} \int_\Omega \{T_{i+1}(x) - F_i(\widehat{\theta}_{f_n}, x)\} \dot{F}_i(\widehat{\theta}_{f_n}, x) \, dx,$$

where $\dot{F}_i(\theta, x) = \partial F_i(\theta, x)/\partial\theta$. In the positive sign region,

$$\dot{F}_{i,1}(\theta, x) = T_i\{C_{\alpha_2}(T_{i-1})(x)\} - C_{\alpha_2}(T_{i-1})(x),$$

and

$$\dot{F}_{i,2}(\theta, x) = C_{\alpha_1}(T_i)'\{C_{\alpha_2}(T_{i-1})(x)\}\{T_{i-1}(x) - x\}.$$

Let

$$s_i(\theta) = \int_\Omega \{T_{i+1}(x) - F_i(\theta, x)\} \dot{F}_i(\theta, x) \, dx, \qquad H = -\mathbb{E}\left\{\frac{\partial s_i(\theta_0)}{\partial \theta^\top}\right\}.$$

Under standard smoothness conditions and nonsingularity of H, a Taylor expansion of the estimating equation gives

$$\sqrt{f_n}(\widehat{\theta}_{f_n} - \theta_0) = H^{-1} \frac{1}{\sqrt{f_n}} \sum_{i=2}^{f_n - 1} \int_\Omega \{T_{i+1}(x) - F_i(\theta_0, x)\} \dot{F}_i(\theta_0, x)\, dx + o_p(1).$$

Thus, in contrast to ATM(1), the leading estimation effect is two-dimensional.

Now consider the ATM(2) residual map

$$R_i(\theta, x) = T_i \circ C_{\alpha_2}(T_{i-2})^{-1} \circ C_{\alpha_1}(T_{i-1})^{-1}(x).$$

At the true value, $R_i(\theta_0, x) = T_{\epsilon_i}(x)$. Writing

$$U_i(\theta, x) = R_i(\theta, x) - x, \qquad \eta_i(x) = T_{\epsilon_i}(x) - x,$$

we have the first-order expansion

$$U_i(\widehat{\theta}_{f_n}, x) = \eta_i(x) + \dot{R}_i(\theta_0, x)^\top (\widehat{\theta}_{f_n} - \theta_0) + o_p(\|\widehat{\theta}_{f_n} - \theta_0\|).$$

Let

$$A_{i-1} = C_{\alpha_{1,0}}(T_{i-1}), \qquad B_{i-2} = C_{\alpha_{2,0}}(T_{i-2}),$$
$$w_i(x) = A_{i-1}^{-1}(x), \qquad z_i(x) = B_{i-2}^{-1}\{w_i(x)\}.$$

A direct chain-rule calculation gives

$$\dot{R}_{i,1}(\theta_0, x) = -T'_{\epsilon_i}(x)\{T_{i-1}(w_i(x)) - w_i(x)\},$$

and

$$\dot{R}_{i,2}(\theta_0, x) = -T'_{\epsilon_i}(x) A'_{i-1}(w_i(x))\{T_{i-2}(z_i(x)) - z_i(x)\}.$$

Consequently, the first-order perturbation of the residual map is not driven only by the current innovation. It also contains the lagged state variables T_{i-1} and T_{i-2}, which themselves depend on past innovations such as $T_{\epsilon_{i-1}}$ and $T_{\epsilon_{i-2}}$.

For the residual ACF, this yields

$$\sqrt{l_n}\widehat{\rho}_{l_n}^{(2)}(k) = \sigma_1^{-2} \left[\frac{1}{\sqrt{l_n}} \sum_i \int_\Omega \eta_i(x)\eta_{i+k}(x)\, dx + \sqrt{l_n} A_k^{(2)\top} (\widehat{\theta}_{f_n} - \theta_0) \right] + o_p(1),$$

where

$$A_k^{(2)} = \mathbb{E} \int_\Omega \{\dot{R}_i(\theta_0, x)\eta_{i+k}(x) + \dot{R}_{i+k}(\theta_0, x)\eta_i(x)\}\, dx.$$

The second term above is the first-order estimation effect. In the ATM(1) case, the corresponding sample-splitting calibration relies on a cancellation identity between the ACF sensitivity and the influence function of the estimator. For ATM(2), the analogous identity would be a matrix condition of the form

$$\Gamma^{(2)} = -A^{(2)}\Omega_\theta,$$

where $A^{(2)} = (A_1^{(2)}, \ldots, A_K^{(2)})^\top$ is the residual-ACF sensitivity matrix, Ω_θ is the covariance matrix of the ATM(2) estimator influence function, and $\Gamma^{(2)}$ is the overlap covariance between the oracle ACF summand and the estimator influence function. This identity is not expected to hold automatically because the nested contraction and inverse-composition operations introduce additional lagged-innovation terms into $\dot{R}_i(\theta_0, x)$.

Therefore, unlike in the first-order ATM setting, the choice $f_n = n/2$ and $l_n = n$ does not by itself guarantee that the ATM(2) estimation effect is removed. The statistic $Q_n^{S,(2)}$ should thus be interpreted as a natural higher-order extension of the ATM(1) splitting statistic, whose calibration requires separate finite-sample assessment or an additional higher-order covariance correction.

To illustrate the preceding discussion, we compute a numerical diagnostic for the ATM(2) matrix cancellation identity. For each Monte Carlo replication, let

$$R = \sqrt{m}\,\widehat{\rho}^{(2)}_{\text{oracle}}, \qquad B = \sqrt{f_n}\,(\widehat{\theta}_{f_n} - \theta_0), \qquad S = \sqrt{m}\{\widehat{\rho}^{(2)}(\widehat{\theta}_{f_n}) - \widehat{\rho}^{(2)}_{\text{oracle}}\}.$$

Here $\theta_0 = (\alpha_{1,0}, \alpha_{2,0})^\top$. We estimate the effective sensitivity matrix $A^{(2)}$ from the regression $S \approx A^{(2)} B$, and compute

$$\widehat{D}^{(2)} = \widehat{\mathrm{Cov}}(R, B) + \widehat{A}^{(2)}\widehat{\Omega}_\theta.$$

If the ATM(2) analogue of the ATM(1) cancellation identity were valid, the matrix $\widehat{D}^{(2)}$ would be close to zero. Table S.1 reports the resulting discrepancy.

Table S.1: Empirical discrepancy for the ATM(2) matrix cancellation identity. The diagnostic uses $n = 800$, $K = 9$, and 500 Monte Carlo repetitions. Here $\widehat{C}_{RB} = \widehat{\mathrm{Cov}}(R, B)$.

(α_1, α_2)	Reps	$\|\widehat{C}_{RB}\|_F$	$\|\widehat{A}^{(2)}\widehat{\Omega}_\theta\|_F$	$\|\widehat{D}^{(2)}\|_F$	Rel.
$(0.2, 0.1)$	500	0.148	0.303	0.158	0.351
$(0.5, 0.2)$	500	0.121	0.196	0.080	0.252
$(-0.2, -0.1)$	500	0.120	0.202	0.088	0.275
$(-0.5, -0.2)$	500	0.106	0.689	0.625	0.786

The discrepancy is not negligible across the four ATM(2) settings. This is in sharp contrast to the ATM(1) numerical validation in Table 1, where the corresponding discrepancy is close to zero. The diagnostic therefore supports the conclusion that the ATM(1) cancellation mechanism does not automatically extend to ATM(2).

A formal ATM(2) analogue of the sample-splitting theorem in the main text would require a separate treatment of the two-dimensional estimation effect and the associated covariance correction terms, and is not pursued here. We therefore view (S.15) as a practically motivated higher-order extension requiring separate calibration. The ATM(2) alternatives in Section S.4 instead evaluate the power of ATM(1)-based diagnostics against omitted second-order dependence.

S.4 Additional Simulation

This section reports additional simulation results under smoother innovation maps

$$e_2(x) = x + Y_2(x^2 - x), \qquad e_3(x) = x + Y_3 x^2 (1-x)^2,$$

where Y_2, Y_3 are generated from the uniform distribution on $[-1, 1]$. All simulation settings are the same as in Section 5 except for the distributions of innovation maps.

Table S.2: ATM(1) size under the power innovation map $e_2(x) = x + Y_2(x^2 - x)$. Entries are empirical rejection rates at the 5% nominal level based on 1000 Monte Carlo repetitions.

K	Method	$\alpha = 0.2$				$\alpha = 0.5$			
		$n = 100$	$n = 200$	$n = 500$	$n = 1000$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
3	Q_n	0.040	0.049	0.045	0.043	0.056	0.065	0.061	0.078
	Q_n^S	0.042	0.057	0.044	0.051	0.062	0.056	0.043	0.049
6	Q_n	0.025	0.033	0.035	0.052	0.039	0.051	0.061	0.051
	Q_n^S	0.043	0.053	0.043	0.060	0.043	0.055	0.050	0.060
9	Q_n	0.018	0.029	0.032	0.042	0.041	0.039	0.055	0.053
	Q_n^S	0.037	0.057	0.055	0.061	0.041	0.049	0.048	0.057
K	Method	$\alpha = -0.2$				$\alpha = -0.4$			
		$n = 100$	$n = 200$	$n = 500$	$n = 1000$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
3	Q_n	0.057	0.050	0.046	0.046	0.057	0.043	0.044	0.043
	Q_n^S	0.051	0.050	0.039	0.043	0.048	0.049	0.047	0.044
6	Q_n	0.033	0.034	0.040	0.063	0.042	0.040	0.053	0.039
	Q_n^S	0.047	0.051	0.048	0.050	0.050	0.049	0.054	0.051
9	Q_n	0.030	0.023	0.029	0.048	0.031	0.037	0.054	0.058
	Q_n^S	0.056	0.055	0.049	0.057	0.051	0.052	0.061	0.053

Tables S.2–S.3 show that the empirical sizes of both diagnostics remain reasonably close to the nominal level. Indeed, the two smoother innovations e_2 and e_3 are generally better than e_1 in size control. Tables S.4–S.5 show that the power against ATM(2) underfitting increases with the sample size and with the magnitude of the omitted second-order dependence. The two innovations e_2 and e_3 yield moderate power for weak alternatives and small sample sizes. The results above showcase the robustness of the proposed methods against different types of innovation maps.

S.5 Additional Empirical Prediction Checks

As an additional visual check of the fitted ATM(1) dynamics, we estimate the fitted distribution for the year 2018 for each airport; see Figure S.1. The one-step-ahead 2018 comparison shows the largest discrepancy for Alaska, while the New York and San Francisco fitted curves are closer to their observed distributions.

For each target year from 2010 to 2018, we also fit the ATM(1) model using the preceding 50 years of annual distributions and compute the Wasserstein distance between the one-step-ahead predicted distribution and the observed distribution. Thus, the 2010 forecast is trained

Table S.3: ATM(1) size under the polynomial innovation map $e_3(x) = x + Y_3 x^2(1-x)^2$. Entries are empirical rejection rates at the 5% nominal level based on 1000 Monte Carlo repetitions.

K	Method	$\alpha = 0.2$				$\alpha = 0.5$			
		$n = 100$	$n = 200$	$n = 500$	$n = 1000$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
3	Q_n	0.047	0.046	0.039	0.040	0.050	0.060	0.055	0.071
	Q_n^S	0.045	0.053	0.037	0.055	0.062	0.053	0.044	0.046
6	Q_n	0.025	0.034	0.032	0.050	0.044	0.048	0.054	0.048
	Q_n^S	0.050	0.052	0.042	0.059	0.056	0.060	0.048	0.063
9	Q_n	0.017	0.025	0.035	0.043	0.042	0.043	0.044	0.053
	Q_n^S	0.037	0.053	0.051	0.061	0.041	0.053	0.047	0.063
K	Method	$\alpha = -0.2$				$\alpha = -0.4$			
		$n = 100$	$n = 200$	$n = 500$	$n = 1000$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
3	Q_n	0.047	0.036	0.050	0.044	0.056	0.043	0.037	0.042
	Q_n^S	0.053	0.050	0.040	0.050	0.054	0.054	0.045	0.044
6	Q_n	0.035	0.036	0.039	0.042	0.046	0.043	0.049	0.044
	Q_n^S	0.045	0.054	0.045	0.049	0.053	0.051	0.053	0.049
9	Q_n	0.029	0.026	0.032	0.037	0.033	0.039	0.052	0.050
	Q_n^S	0.054	0.058	0.048	0.055	0.050	0.055	0.060	0.051

Table S.4: ATM(2) underfitting power under the power innovation map $e_2(x) = x + Y_2(x^2 - x)$. The data are generated from ATM(2), while an ATM(1) model is deliberately fitted. Entries are empirical rejection rates at the 5% nominal level based on 1000 Monte Carlo repetitions.

K	Method	$(\alpha_1, \alpha_2) = (0.2, 0.1)$				$(\alpha_1, \alpha_2) = (0.5, 0.2)$			
		$n = 100$	$n = 200$	$n = 500$	$n = 1000$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
3	Q_n	0.074	0.114	0.298	0.634	0.290	0.577	0.949	1.000
	Q_n^S	0.119	0.183	0.429	0.768	0.326	0.604	0.950	1.000
6	Q_n	0.044	0.074	0.230	0.491	0.187	0.410	0.822	0.941
	Q_n^S	0.109	0.161	0.340	0.643	0.288	0.526	0.939	0.999
9	Q_n	0.027	0.050	0.176	0.424	0.162	0.338	0.769	0.957
	Q_n^S	0.102	0.136	0.300	0.569	0.268	0.458	0.896	0.998
K	Method	$(\alpha_1, \alpha_2) = (-0.2, -0.1)$				$(\alpha_1, \alpha_2) = (-0.5, -0.2)$			
		$n = 100$	$n = 200$	$n = 500$	$n = 1000$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
3	Q_n	0.086	0.141	0.328	0.601	0.249	0.532	0.909	0.979
	Q_n^S	0.098	0.161	0.394	0.716	0.272	0.593	0.950	0.999
6	Q_n	0.069	0.103	0.234	0.468	0.199	0.388	0.841	0.968
	Q_n^S	0.083	0.119	0.290	0.571	0.232	0.461	0.907	0.997
9	Q_n	0.047	0.076	0.186	0.388	0.169	0.335	0.775	0.968
	Q_n^S	0.080	0.104	0.251	0.508	0.208	0.396	0.858	0.995

Table S.5: ATM(2) underfitting power under the polynomial innovation map $e_3(x) = x + Y_3 x^2(1-x)^2$. The data are generated from ATM(2), while an ATM(1) model is deliberately fitted. Entries are empirical rejection rates at the 5% nominal level based on 1000 Monte Carlo repetitions.

K	Method	$(\alpha_1, \alpha_2) = (0.2, 0.1)$				$(\alpha_1, \alpha_2) = (0.5, 0.2)$			
		$n = 100$	$n = 200$	$n = 500$	$n = 1000$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
3	Q_n	0.075	0.108	0.294	0.651	0.290	0.596	0.960	1.000
	Q_n^S	0.126	0.170	0.430	0.763	0.347	0.609	0.951	1.000
6	Q_n	0.050	0.069	0.229	0.504	0.180	0.423	0.861	0.974
	Q_n^S	0.120	0.146	0.342	0.651	0.312	0.537	0.936	1.000
9	Q_n	0.031	0.050	0.185	0.415	0.162	0.349	0.803	0.977
	Q_n^S	0.108	0.123	0.305	0.560	0.281	0.465	0.887	1.000
K	Method	$(\alpha_1, \alpha_2) = (-0.2, -0.1)$				$(\alpha_1, \alpha_2) = (-0.5, -0.2)$			
		$n = 100$	$n = 200$	$n = 500$	$n = 1000$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
3	Q_n	0.083	0.143	0.357	0.658	0.282	0.579	0.943	0.984
	Q_n^S	0.101	0.180	0.435	0.743	0.307	0.638	0.972	1.000
6	Q_n	0.065	0.097	0.245	0.532	0.215	0.435	0.898	0.976
	Q_n^S	0.084	0.133	0.328	0.629	0.256	0.525	0.942	0.999
9	Q_n	0.046	0.075	0.195	0.459	0.175	0.376	0.832	0.983
	Q_n^S	0.085	0.111	0.261	0.556	0.219	0.443	0.895	0.996

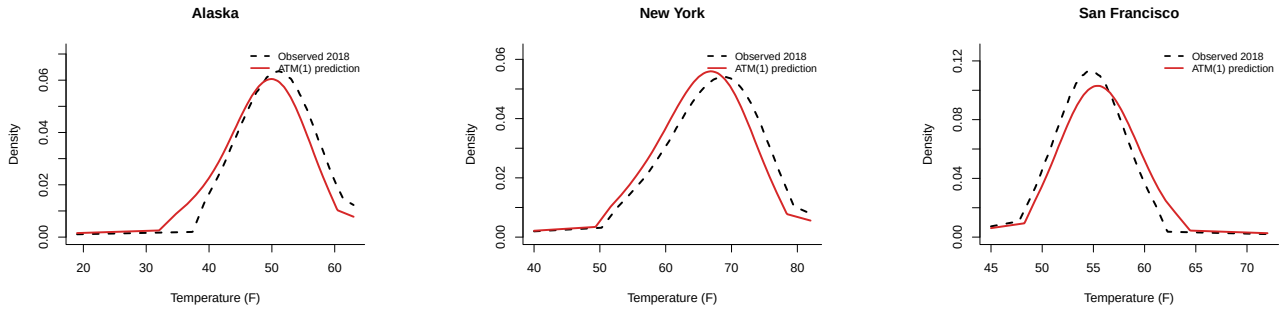


Figure S.1: Model validation for the year 2018 using the corrected fitted ATM(1) model. The observed distribution of daily minimum summer temperatures is shown in black, compared against the predicted distribution from the fitted ATM(1) model (red) for each airport.

on 1960–2009 and the 2018 forecast is trained on 1968–2017. The results are summarized in Table S.6.

Table S.6: Wasserstein distances between predicted and observed distributions, computed with the corrected fitted ATM(1) model.

Year	**Alaska**	**New York**	**San Francisco**
2010	1.0762	2.7729	1.4892
2011	1.0006	1.3680	0.6053
2012	0.6549	0.9243	0.7992
2013	2.4319	0.9404	1.6104
2014	1.3588	1.3556	3.9231
2015	1.6600	2.7617	2.2344
2016	3.2928	2.0377	0.5082
2017	1.9597	0.3613	2.4170
2018	2.0724	1.4287	1.0170
average	1.7230	1.5501	1.6226

Alaska exhibits the largest average prediction error, whereas New York has the smallest average error, with San Francisco lying in between. This forecasting comparison is broadly consistent with the diagnostic results in the main text.

S.6 Numerical Verification at Different Lag Orders

We examine condition (4.13) at $K = 3, 6, 9, 12$ for the three innovation maps and four parameter values in Table 1 of the main text. The simulation settings are unchanged: each series contains $n = 5000$ observations after a burn-in of 300, and the maps are evaluated on 601 equally spaced grid points on $[0, 1]$. Each setting uses 1000 Monte Carlo repetitions.

Table S.7 reports the resulting means and standard deviations; the $K = 12$ entries reproduce Table 1 up to rounding. The mean discrepancies remain on comparable scales across the four lag choices within each innovation family, although their values depend on K and α_0. Since both measures average over the first K lags, they need not decrease or increase uniformly as K grows. The numerical evidence reported in the main text is therefore not specific to $K = 12$; it remains evidence for the configurations studied, rather than a proof of the cancellation identity for general ATM models.

References

Davis, R. A. and Fernandes, L. (2025). Sample splitting and assessing goodness-of-fit of time series. *Biometrika*, 112(2):Paper No. asaf017, 21.

Wu, W. B. and Shao, X. (2004). Limit theorems for iterated random functions. *Journal of Applied Probability*, 41(2):425–436.

Zhu, C. and Müller, H.-G. (2023). Autoregressive optimal transport models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):1012–1033.

Table S.7: Numerical sensitivity of condition (4.13) to K, with $n = 5000$ and 1000 Monte Carlo repetitions. Each entry is mean (standard deviation). All entries for e_1 and e_3 are in units of 10^{-6}; those for e_2 are in units of 10^{-4}.

Function	K	Difference	$\alpha_0 = 0.5$	$\alpha_0 = 0.2$	$\alpha_0 = -0.2$	$\alpha_0 = -0.4$
$e_1(x)$	3	L^1	3.765(0.796)	1.965(0.818)	4.792(0.837)	3.080(0.809)
		L^2	4.465(0.877)	2.245(0.955)	5.984(0.667)	3.560(0.911)
	6	L^1	2.703(0.550)	1.933(0.591)	3.311(0.601)	2.457(0.570)
		L^2	3.474(0.629)	2.319(0.713)	4.526(0.533)	2.986(0.654)
	9	L^1	2.336(0.449)	1.906(0.488)	2.817(0.504)	2.247(0.486)
		L^2	3.066(0.529)	2.322(0.592)	3.924(0.465)	2.770(0.564)
	12	L^1	2.140(0.385)	1.892(0.442)	2.577(0.443)	2.138(0.420)
		L^2	2.827(0.461)	2.326(0.530)	3.593(0.431)	2.653(0.489)
$e_2(x)$	3	L^1	2.322(0.262)	0.971(0.385)	1.876(0.360)	0.837(0.372)
		L^2	2.584(0.183)	1.117(0.466)	2.281(0.239)	0.947(0.412)
	6	L^1	1.673(0.239)	1.030(0.313)	1.474(0.294)	0.935(0.282)
		L^2	2.036(0.174)	1.237(0.376)	1.865(0.239)	1.120(0.337)
	9	L^1	1.431(0.211)	1.057(0.270)	1.342(0.251)	0.966(0.239)
		L^2	1.796(0.168)	1.286(0.325)	1.708(0.225)	1.178(0.287)
	12	L^1	1.314(0.188)	1.071(0.231)	1.277(0.223)	0.984(0.215)
		L^2	1.666(0.162)	1.314(0.280)	1.625(0.212)	1.207(0.256)
$e_3(x)$	3	L^1	3.327(1.569)	3.756(2.023)	4.018(1.975)	3.393(1.735)
		L^2	3.887(1.807)	4.658(2.447)	4.833(2.436)	4.036(2.052)
	6	L^1	3.960(1.268)	4.488(1.534)	4.645(1.548)	4.102(1.332)
		L^2	4.834(1.514)	5.649(1.842)	5.737(1.870)	5.087(1.627)
	9	L^1	4.178(1.074)	4.734(1.262)	4.788(1.288)	4.363(1.121)
		L^2	5.160(1.282)	5.952(1.512)	5.938(1.513)	5.437(1.355)
	12	L^1	4.283(0.963)	4.900(1.146)	4.885(1.116)	4.491(1.008)
		L^2	5.312(1.132)	6.153(1.363)	6.076(1.317)	5.603(1.203)