Path-Coupled Bellman Flows for Distributional Reinforcement Learning

Boyang Xu, Qing Zou, Siqin Yang, Hao Yan

School of Computing & Augmented Intelligence (SCAI), Arizona State University, Tempe, AZ, USA

GitHub arXiv

Motivation

Flow matching requires the generation process to start from a fixed simple prior (e.g., Gaussian), but the distributional Bellman update inherently violates this.

$$Z^{\pi}(s,a) = R(s,a) + \gamma \cdot Z^{\pi}(S', A')$$

What we solved

1. **Boundary mismatch.** Flow matching requires $t = 0$ to be the simple prior X_0. But the uncorrected Bellman path starts at $R + \gamma X_0 \neq X_0$ - shifted by the reward and discounted successor value, so the flow matching objective is ill-posed at $t = 0$.
2. **Independent noise.** When current and successor flows draw independent noise, their intermediate trajectories are unrelated, so Bellman consistency can only be enforced at the endpoint $t = 1$, yielding high-variance, destabilizing targets.
3. **Variance Control.** A λ-parameterized control variate on the coupled paths reduces target variance without breaking Bellman consistency.

Method: Path-Coupled Bellman Flows (PCBF)

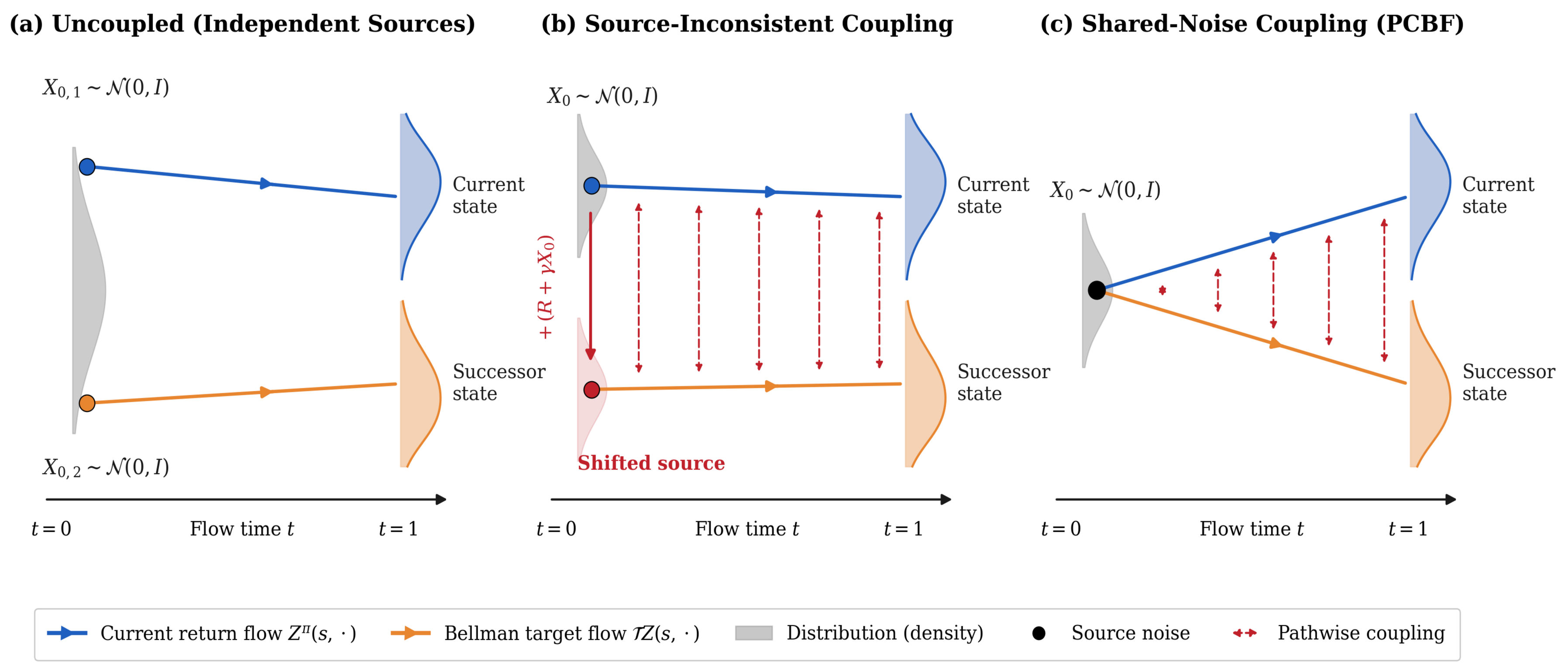

Fig 1. Path-coupled Bellman geometry.

Successor path $Z_t^{s'} = (1-t)\cdot X_0 + t\cdot X',\ X_0 \sim \mathcal{N}(0,1)$

Current path $Z_t^{s} = (1-t)\cdot X_0 + t\cdot (R + \gamma X')$

Training target $u_t^{\lambda} = (R + \gamma X' - X_0) + \lambda \cdot [v_{\theta^-}(t, Z_t^{s'} | s', a') - (X' - X_0)]$

Results on MRPs

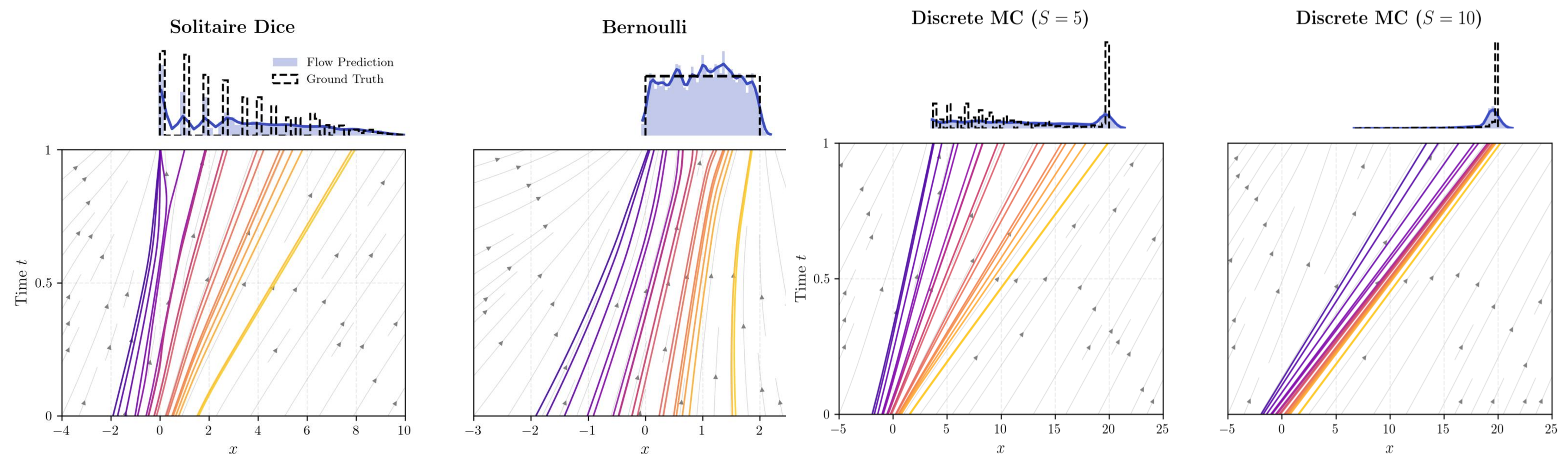

Fig 2. PCBF recovers ground-truth return laws (dashed) on Solitaire, Bernoulli, and Discrete MC MRPs.

Results on Offline RL

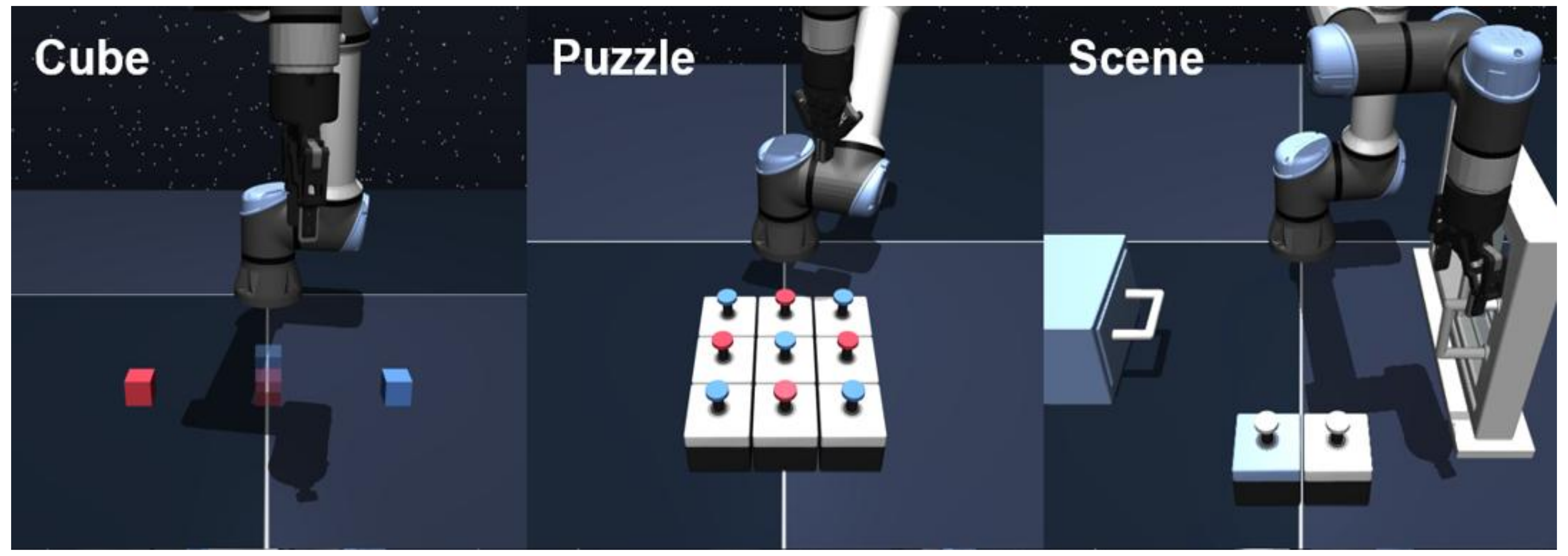

Fig 3. Cube/Puzzle/Scene manipulation tasks from OGBench.

Table 1. Offline RL Results: we report performance on OGBench and D4RL. Results are averaged over 8 random seeds (4 seeds for pixel-based tasks).

Domain	IQN	CODAC	FloQ	FQL	IQL	Value Flows	PCBF(Ours)
cube-double-play	42 ± 8	61 ± 6	47 ± 14	29 ± 6	7 ± 1	69 ± 4	**71 ± 5**
scene-play	40 ± 1	55 ± 1	**58 ± 4**	56 ± 2	28 ± 3	**59 ± 4**	54 ± 4
puzzle-4x4-play	27 ± 4	20 ± 18	28 ± 6	17 ± 5	7 ± 2	27 ± 4	**30 ± 4**
cube-triple-play	6 ± 0	2 ± 1	8 ± 3	4 ± 2	1 ± 1	**14 ± 3**	4 ± 1
D4RL Adroit	66 ± 5	69 ± 0	70 ± 5	**71 ± 4**	70	65 ± 2	**69 ± 2**
visual-antmaze-teleport	4 ± 2	-	-	5 ± 2	6 ± 3	13 ± 4	**14 ± 4**
visual-cube-double-play	1 ± 0	-	-	6± 1	11 ± 6	**13 ± 2**	3 ± 0

Additional Results

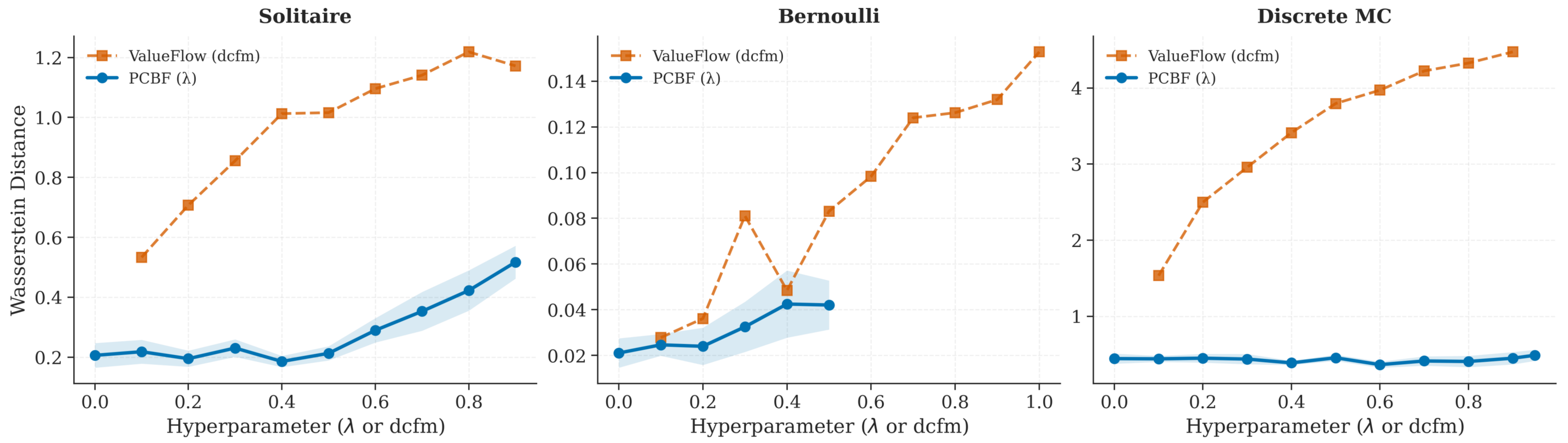

Fig 4. PCBF stays robust across λ; Value Flows degrades as DCFM grows.

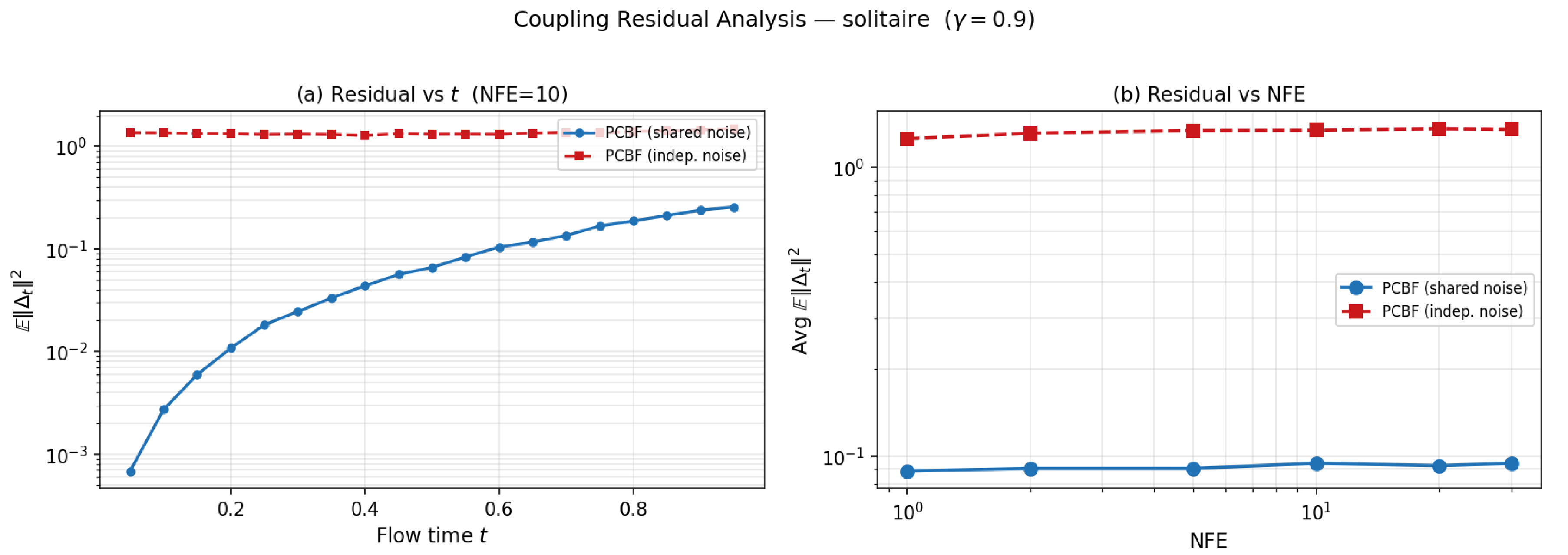

Fig 5. Shared-noise coupling yields smaller Bellman residual across flow-time and N function evaluations (NFE).

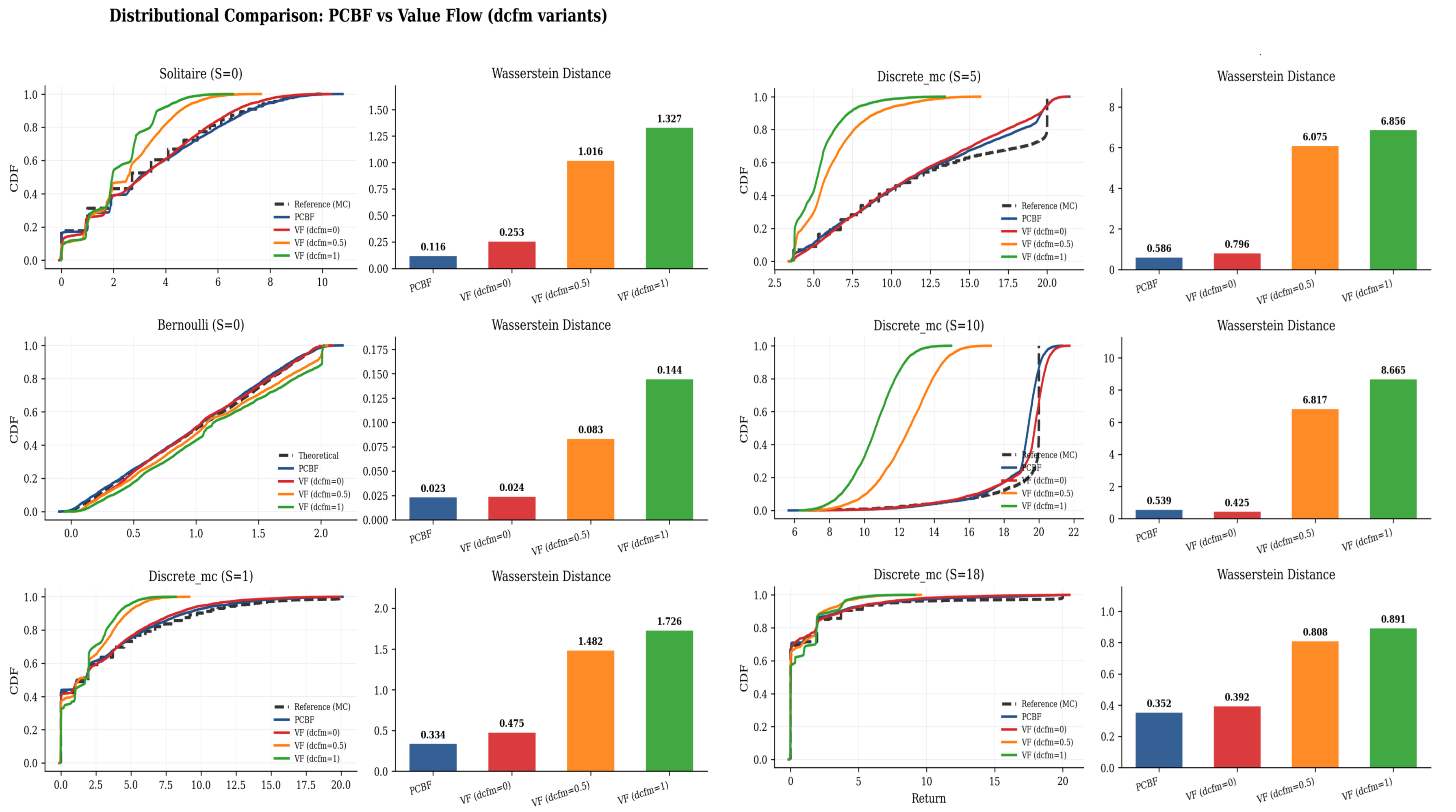

Fig 6. PCBF (blue) tracks ground-truth CDFs; VF underestimates variance at high DCFM.